DOCUMENT RESUME

ED 300 417 TM 012 415

AUTHOR Wilson, Leslie R.; And Others
TITLE Determining Validity and Reliability of Locally Developed Assessments--1988.
INSTITUTION Homewood-Flossmoor Community High School District 233, Flossmoor, IL.
SPONS AGENCY Illinois State Board of Education, Springfield. Dept. of School Improvement Services.
PUB DATE Apr 88
NOTE 129p.
PUB TYPE Guides - Non-Classroom Use (055)

EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS Academic Achievement; *Achievement Tests; *Criterion Referenced Tests; Educational Assessment; Educational Change; Educational Improvement; Elementary Secondary Education; Learning Processes; Teacher Developed Materials; *Teacher Made Tests; Test Construction; Test Reliability; *Test Validity
IDENTIFIERS *Illinois; Paper and Pencil Tests

ABSTRACT

The three chapters of this handbook are designed to insure that any school can use teacher-made tests and performance assessments as components of a learning assessment plan. Focus is on criterion-referenced achievement tests developed in Illinois for assessing the cognitive and psychomotor domains of learning. The initial chapter is designed for classroom teachers and administrators in Illinois to translate the Illinois reform legislation on learning objectives and local systems into local practice. Good practices for paper-and-pencil testing are reviewed. The second chapter describes various methods of checking reliability and gives examples of each. The final chapter, on performance assessment, describes sample assignments and evaluative criteria for some performance areas, with considerations for scoring and a set of sample rating scales. These performance instruments are illustrative only. A locally made companion videotape was produced for use in local rater-training workshops. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED300417

VALIDITY RELIABILITY

Determining Validity and Reliability of Locally Developed Assessments — 1988 —

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

D. G. Gill

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 012 415

Development of this manual has been funded by the
Illinois State Board of Education
Department of School Improvement Services

BEST COPY AVAILABLE

DETERMINING VALIDITY AND RELIABILITY OF LOCALLY DEVELOPED ASSESSMENTS 1988

Leslie R. Wilson, Project Director
Barbara C. Scherbarth, Assistant Project Director
Henry M. Brickell
Samuel T. Mayo
Regina H. Paul

Homewood-Flossmoor Community High School
District 233

© 1988 Illinois State Board of Education

FOREWORD

Through funding provided by the Illinois State Board of Education, Department of School Improvement Services and with the full support and cooperation of the Homewood-Flossmoor Community High School District 233 Board of Education and Administration, we were able to assemble a project team which produced this manual and a training videotape.

We became involved because we wanted to guarantee that any school including our own could continue to use teacher-made tests and performance assessment as components of its Learning Assessment Plan. Our intent was to make validity and reliability measures accessible to all Illinois teachers. We hoped to provide an intermediate ground between the standard texts in the field and everyday classroom practice in measurement.

Because of their expertise and experience, we asked Henry M. Brickell, Samuel T. Mayo, and Regina H. Paul to work with us as project consultants. Henry M. Brickell is president and Regina H. Paul is research associate of Policy Studies in Education, 310 East Shore Road, Great Neck, New York 11023, (516) 829-4144. Policy Studies in Education is a nonprofit educational research and development organization serving local school districts, state education departments, the federal government, foundations, and state and national professional organizations. Samuel T. Mayo is Professor Emeritus, Loyola University of Chicago; Fellow, American Psychological Association; former Book Review Editor of Psychometrika; and consultant to school districts, professional schools, publishers, and businesses. His consulting service, 2525 Park Lane, Glenview, Illinois 60025, (312) 724-1529, is for educational and psychological measurement, research, and evaluation. The five of us are responsible for the content of this document.

Determining Validity and Reliability of Locally Developed Assessments contains three chapters: "Making Sure," "Ensuring Reliability," and "Performance Assessment." The initial chapter, "Making Sure," is intended for classroom teachers and administrators in Illinois to translate the Illinois reform legislation on learning objectives and assessment systems into local practice. It includes general introductory sections on objectives, paper and pencil assessment, and performance assessment.

The chapter was reviewed by twenty-seven K-12 teachers and administrators before its completion. "Ensuring Reliability" describes various methods of checking reliability and examples of each method. "Performance Assessment" contains a description of sample assignments and evaluative criteria of five performance areas, considerations for scoring, and a set of sample rating scales. We did not systematically estimate the validity and reliability of the performance instruments. The performance instruments are illustrative only; we have not validated them and are not specifically recommending or endorsing particular forms.

In addition to this manual we have produced a companion videotape for use in rater-training workshops. The videotape contains regular high school students throwing a sixteen inch softball, shooting a lay-up, executing a dance combination, acting as an older person, and delivering a brief speech. For this portion of the project we assembled video production and curricular teams from among the staff and students at Homewood-Flossmoor Community High School. We also called upon H-F staff members for materials processing, layout, and assembly.

We feel we have accomplished what we set out to do: provide an accessible route to establishing validity and reliability on teacher-made tests, and a vehicle for incorporating performance assessment in local assessment plans. We sincerely hope you find this manual useful as you respond to the Illinois reform package.

Leslie R. Wilson
Barbara C. Scherbarth
April 1988

TABLE OF CONTENTS

Page

Foreword i

Introduction vii

CHAPTER ONE: MAKING SURE

Establishing Good Objectives 1

Why Is It So Important to Establish Good Objectives? 1
What Does a Good Objective Look Like? 1
How to Check How Good Your Objectives Are 4

Establishing Good Paper-and-Pencil Testing 5

Establishing Good Questions for Paper-and-Pencil Testing 6

How Many Good Questions 7

Establishing Validity of Paper-and-Pencil Testing 8

What Is Content Validity? 8
How to Establish Content Validity 9

Establishing Reliability of Paper-and-Pencil Testing 10

What Is Reliability? 10
Preparing to Determine Consistency 11
How to Calculate Consistency 12
How Much Consistency Is Enough? 13

Establishing Good Performance or Product Assessments 15

Assessing the Performance or the Product--or Both 15
Recording the Performance Before Assessing It 16
Building Archives of Performances and Products 16
Measuring Performances 17
Considerations about Judging 17

Establishing Good Rating Scales 18

What Will the Scale Ask the Teacher to Judge? 19
How Many Points Will the Scale Have? 19
How Will the Points Be Labelled? 20
How Many Good Rating Scales? 21

Establishing Validity of Performance or Product Assessments 22

How to Establish Content Validity 22

Establishing Reliability of Performance or Product Assessments 23

Training the Teachers to Use Rating Scales 24
Calculating Reliability 24
Reliability of Ratings by One Teacher 24
Reliability of Ratings by Two or More Teachers 26

CHAPTER TWO: ENSURING RELIABILITY

Establishing Reliability for Paper and Pencil Testing 31

Theoretical Considerations 31

Consistency Methods 32

Internal Consistency Reliability 32

Split-Half Reliability for a Test ("Long" Method) 34

Reliability Index 36

Split-Halves: Mastery Consistency Methods 38

Consistency Indexes for Criterion-Referenced (Mastery) Tests 40

Establishing Reliability of Performance Ratings 43

Performance Conditions 43

Selecting Raters for Performance Ratings 43

Training Raters for Performance Ratings 43

Common Errors in Rating 44

Single-Rater Reliability 46

Two-Rater Reliability 47

Multi-Rater Reliability 49

Worksheet 53

Chapter Bibliography 54

CHAPTER THREE: PERFORMANCE ASSESSMENT

Performance Assessment: Objectives and Assignments 59

Characteristics of Good Assignments 59

Process for Designing Assignments 60

Examples 60

Rating Performance 66

Choosing the Right Criteria for Judging Performance 66

Example Criteria 67

Rating Scales, Checklists, and Scoring 70

Holistic vs. Analytic Scoring 70

Directory for Sample Rating Forms 72

Sample Rating Forms 75

Art 75
Dance 76
Drama 80
Music 85
Speech 97
Composition 102

Developing Inter-Rater Reliability: Issues to Consider 103

Standard Score 103

Standard Deduction 104

Recording Performance Ratings 105

Student Profiles 105

Objective Profiles 106

Class Roster Profiles 106

Chapter Bibliography 108

PROJECT NOTES 110

ANNOTATED BIBLIOGRAPHY 115

Special Thanks 119

INTRODUCTION

This booklet is written to help teachers and administrators in local school districts in Illinois establish validity and reliability for their own criterion-referenced tests. Such tests, written and designed locally, may be used to measure the achievement of student learning objectives in accordance with a district's learning assessment plan.

This booklet is about:

Locally-designed tests only --that is, tests designed by one teacher for one classroom, several teachers for one school, or a group of teachers for the whole school district. This booklet is not about state tests, national tests, or commercially-published tests.

Criterion-referenced tests only --that is, tests written to measure the achievement of a given set of specific objectives, with the standard set by the teacher, school, or district. This booklet is not about norm-referenced tests, with the standard set by the achievement of other students who take the tests.

Achievement tests only --that is, tests that measure the acquisition of knowledge or skills described by a given set of specific objectives. This booklet is not about ability tests or aptitude tests.

The cognitive and psychomotor domains of learning only --that is, knowledge/thinking and physical skills. This booklet is not about the validity and reliability of testing the affective domain of learning (attitudes, values, beliefs, and feelings).

This booklet is written for readers without a background in statistics or measurement. Thus, it does not offer explanations of statistical techniques. And it does not offer a variety of alternative statistical procedures. It rather gives only the most straightforward procedures for proving the validity and reliability of criterion-referenced tests in local school districts.

This booklet may be used by any classroom teacher of any subject and any grade (as well as by any administrator) to help meet Illinois' high standards for locally-designed, criterion-referenced, achievement tests of cognitive and psychomotor skills.

MAKING SURE

CHAPTER ONE

ESTABLISHING GOOD OBJECTIVES

Why Is It So Important To Establish Good Objectives?

Advice about valid and reliable testing could scarcely begin with any other point. Objectives start the testing machinery ticking. Validity is the content match between an assessment and the objective. Does the test item cover the same materials as the objective? Reliability is the extent to which a test can be depended upon to provide consistent, nearly error-free information. Without clear objectives and focused instruction, there can be no valid tests. There could be reliable tests, but they would not be worth giving.

Objectives specify the student learning that is to be tested. Just as objectives guide your teaching, so they guide your testing. Neither teaching nor testing should wander around without the guidance of good, clear objectives--clear to you, your students and their parents, your principal, your superintendent, and your board of education.

What Do Good Objectives Look Like?

They have several characteristics:

1. Objectives describe what students will know, feel, or be able to do. They describe what the students will learn, not what you will do as a teacher to cause them to learn. Some objectives describe knowledge and thinking skills (the cognitive domain); some describe feelings, attitudes, beliefs, and values (the affective domain); some describe physical skills (the psychomotor domain). You should have both cognitive and affective objectives in your lesson plans, regardless of what subject or grade you teach. You may also have psychomotor objectives in your lesson plans, depending on what subject and grade you teach. (Remember: Psychomotor objectives are not found only in physical education. They include things like using lab equipment in science, using compasses and protractors in mathematics, handwriting in English, drawing maps in social studies, drilling in industrial arts, and slicing in home economics.)

To read further about the three domains of learning, try:

Bloom, Benjamin S., Ed. Taxonomy of Educational Objectives, The Classification of Educational Goals, Handbook I: Cognitive Domain. New York: David McKay Company, Inc. . 1956.

Krathwohl, David R., Benjamin S. Bloom, and Bertram B. Masia. Taxonomy of Educational Objectives, The Classification of Educational Goals, Handbook II: Affective Domain. New York: David McKay Company, Inc. 1964.

Simpson, Elizabeth J. "The Classification of Educational Objectives in the Psychomotor Domain." The Psychomotor Domain. Washington, D.C.: Gryphon House. 1972. Pp. 43-56.

2. Objectives are specific.

They must be specific enough to guide your daily teaching--or the teaching done by a newcomer to your district. They must be specific enough to guide test question writing. If they are too broad, both teaching and testing might miss the intended mark. For example:

Shoots baskets is too broad; *Shoots lay-ups with the dominant hand* is better.

Appreciates paintings is too broad; *Appreciates nineteenth century Impressionists' paintings* is better.

Uses commas is too broad; *Uses commas to separate tems in a series* is better.

You can see that if the objective were "Uses commas," there would be many, many uses you could teach and many, many uses you could test. "Uses commas to separate items in a series" better focuses both your teaching and your testing. That's helpful both to you and your students.

With objectives this specific, how many would you need to cover a year's worth of third grade music , seventh grade math, or eleventh grade science? About 75 per subject per grade in the elementary years and about 125 per subject per grade (that is, per course) in the secondary years. (These are just ballpark estimates. You may find you will want more; you probably should not have fewer.)

3. Objectives are not repeated from grade to grade.

   Because they are specific and because they break down learning into reasonably-sized pieces, objectives should not be repeated from the previous year's work. Of course, you may need to do some review; but "review" is not new teaching. Students should be expected to learn certain knowledge, acquire certain values, and master certain skills and then hold onto that learning as they continue through their school years.

   Repeating objectives verbatim from grade to grade will make high school tests look like elementary school tests, which would be a mistake. As material gets more complex, the ability to interpret properly the material is not repetition but extension and development of previously introduced, more basic material. For example, an objective for students in grade two might be "gets the stated main idea in a paragraph." So, the test question would include one paragraph that could be read by a second grader. By sixth grade the object would be "gets the implied main idea in an essay." So, the test question would include an essay of at least 6-8 paragraphs that would be challenging enough to sixth graders.

4. Objectives are clearly and simply worded.

   They should be clearly enough worded that another teacher who did not write them could read them and teach them without having to ask what they mean. They should be simply enough worded that any community member of the board of education could read them and understand exactly what is being taught in the schools.

   Lack of clarity--just like lack of specificity--makes the writing of valid test questions impossible.

5. Objectives in the cognitive and affective domains begin with verbs that specify precise levels or degrees of thinking or feeling.

   The verb that begins an objective is critically important for guiding both your teaching and your test question writing. Imagine how differently you would teach and test objectives that begin with these verbs:

Cognitive Domain		**Affective Domain**
Defines	*Applies*	*Accepts*
Lists	*Analyzes*	*Enjoys*
Understands	*Compares*	*Values*
Predicts	*Judges*	*Prefers*

Note: Objectives in the psychomotor domain usually begin with verbs that name the skills being performed. For example:

Throws	*Prints*
Measures	*Sews*
Draws	*Kicks*
Swims	*Welds*

The precise level of physical ability called for from the student is often specified as a criterion in the objective. For example:

Throws a softball 90 feet to a partner.
Measures to 1/16 of an inch.

How To Check How Good Your Objectives Are

The writer of a set of objectives is not a good judge of how good those objectives are. A reader of the objectives must judge that. Teacher committees that write objectives could use either of these two ways to find out how good their objectives are:

1. Ask several teachers of the appropriate grade and learning area to read each objective independently and explain to the committee what each means, perhaps by explaining the kinds of lessons they would use to teach it.

2. Ask several teachers of the appropriate grade and subject to read each objective independently and write a test question (or a performance exercise) to test whether students learned it.

Both methods will quickly expose any lack of clarity in the written objectives--whether they were not specific enough, not clear enough, and so on. The second method is a tougher test of how good the objectives are; you should use it.

To read more about writing good objectives, see:

Mager, R.F., Preparing Instructional Objectives, 2nd Edition, Palo Alto, California: Fearon, 1976.

ESTABLISHING GOOD PAPER-AND-PENCIL TESTING

Any teacher knows many types of paper-and-pencil test question formats for objective test questions: multiple choice, true-false, matching, and short answer, to name the most popular. (Essay tests require paper and pencil, of course, but they cannot be objectively measured; therefore, they are included in the next section of this booklet, Establishing Good Performance Assessments.)

Any one of these question formats--or all of them--can be used in locally-designed tests. The methods for establishing validity and reliability discussed in this section do not depend on any certain question format; they will work for all types of questions.

However, in practice teachers and administrators usually choose multiple-choice questions for these reasons:

They are quick, easy, and inexpensive to score.

They allow for many questions to be asked about the same objective without one question giving away the answer to another.

They represent a style familiar to both students and teachers.

They can be used to test both simple knowledge (such as the recall of facts) and complex critical thinking (such as the evaluation of policies).

They give students practice for taking the many other multiple-choice tests they will meet in their school careers (e.g., nationally-standardized tests, college entrance tests).

The only drawback is that good multiple-choice questions are hard to write. Most teachers have not had much formal training in how to write them. While this booklet does not have such training as its primary purpose, it can offer a few tips for you to keep in mind if you decide to use multiple-choice test questions in your locally-designed tests.

Establishing Good Questions for Paper-and-Pencil Testing

Use this checklist when writing and editing your multiple-choice questions (excerpted, with adaptations, from William A. Mehrens and Irvin J. Lehmann's Measurement and Evaluation in Education and Psychology, New York: Holt, Rinehart and Winston, 1978).

1. Does the question match the objective perfectly? (This topic will be discussed in detail in the next section of this booklet.)

2. Has the question been clearly written?

3. Has wordiness in the question been eliminated?

4. Is the main idea in the question--not in the answers?

5. Has the question been written to avoid unnecessary repetition of key words or phrases in each answer?

6. Have the answers been arranged in some systematic fashion, such as alphabetically, numerically, or by length of response?

7. Are all answers reasonable--including all the wrong answers?

8. Have all "giveaway" clues been avoided (e.g., the right answer is the only one that fits the question grammatically; the longest answer is always the right answer; the right answer to one question gives away the right answer to another question)?

9. Are the right answers scattered throughout the questions' options with approximately equal frequency--that is, A is used as often as C for the right answer?

10. Is there only one right answer?

11. Has "all of the above" been avoided?

12. If a negative appears in the question, has the negative always been underlined or written in capital letters?

13. Is the question non-discriminatory to students of all backgrounds--male and female, black and other, rich and poor, native and non-native English speakers?

Many good references are available for helping you write good multiple-choice questions (and good true-false, matching, and short answer questions as well). if you are interested in detailed reading on this subject, try any one of these:

"Bias Issues in Test Development." Amherst, Massachusetts: National Evaluation Systems, Inc. 1987.

Gronlund, Norman E. Constructing Achievement Tests. Englewood Cliffs: Prentice-Hall, Inc. 1982. Pp. 50-59.

Gronlund, Norman E. Preparing Criterion-Referenced Tests for Classroom Instruction. New York: The MacMillan Company. 1973. Pp. 34-41.

Thorndike, Robert L., and Elizabeth P. Hagen. Measurement and Evaluation in Psychology and Education. Fourth Edition. New York: John Wiley & Sons. 1977. Pp. 218-241.

Burry, James, Joan Herman, and Eva L. Baker. "A Practical Approach to Local Test Development." Resource Paper No. 6. Los Angeles: Center for the Study of Evaluation, Graduate School of Education, University of California at Los Angeles. 1984. Pp. 21-28.

Ory, John C. "Improving Your Test Questions." An Assessment Handbook for Illinois Schools. Illinois State Board of Education. 1988.

How Many Good Questions?

If your objectives are specific, as we have said they should be, four good questions will make up a reasonable test of the learning called for by the objective. (Science Research Associates use three questions per objective for their criterion-referenced testing.) In writing four questions--especially if they are multiple-choice questions--you will have to make sure that one question does not give away the answer to another. Remember that the student will see all four questions at the same time.

One note on test construction: If you are testing more than one objective at a time (which you undoubtedly will be doing), do not put all four questions on a given objective together on the test. Rather, place the four questions throughout the test so that they are interspersed with questions on other objectives which prevents one question from giving away the answer to another on the same objective.

Establishing Validity of Paper-and-Pencil Testing

What Is Content Validity?

Now that you have written good objectives and at least four good test questions to match each objective, it is time to establish the validity of those test questions for each objective. To say that a four-question test for an objective has content validity is to say that those four test questions adequately represent the learning called for by that objective.

Let's say that objective #1 is : "Identifies a variety of poetic forms." Suppose the four test questions you wrote to match that objective all asked students to identify Shakespearean sonnets. Is your four-question test content valid? No, because four questions about sonnets do not adequately represent "poetic forms." Did you forget about Petrarchan sonnets? Ballads? Epics? Cinquains? Haiku? To make your four questions adequately represent the learning called for by the objective, you would have to ask students about four different poetic forms. (Of course, you could argue that four questions could not cover all the poetic forms either. In that case, you might want to add some more questions to cover additional poetic forms. Or, you could say--and reasonably so--that questions on four different poetic forms do represent the learning called for by the objective, even if they do not cover every single poetic form imaginable.)

Let's say that objective #2 is : "Compares and contrasts the writing styles of l9th century American novelists." Suppose the four test questions you wrote to match that objective gave students excerpts from four different l9th century American novels and asked students to identify the authors by their style. Is your four-question test valid? No, because four questions about identifying styles do not represent the learning called for by the objective, which is comparing and contrasting styles. To make your four questions match the learning called for by the objective, you would have to ask students to compare the styles of Henry James and Nathaniel Hawthorne, or contrast the styles of Herman Melville and Stephen Crane, and so on.

Let's say that objective #3 is : "Explains the central ideas of well-known essays of l9th century American writers." Suppose the four test questions you wrote to match that objective asked students to explain the central idea of Ralph Waldo Emerson's " Concord Hymn"--a well-known poem. Is your four-question test content valid? No, because the objective called for learning about essays--not poems--even though Emerson himself was a well-known essayist. (You might also note here that four questions on Emerson's well-known essay "Self-Reliance" would have missed the boat, too, because four questions on any one essay could not represent l9th century American essays, as explained earlier.) To make your four questions match the learning called for by the objective, you would have to ask students to explain the central ideas of four different well-known l9th century American essays--if not different essayists.

Content validity, then, has three simple aspects:

1. Do all the questions about an objective taken together represent the full range of the learning called for by the objective (in the example, poetic forms)?

2. Does each separate question ask the student to use the same intellectual process called for by the objective (in the example, comparing and contrasting)?

3. Is each separate question about the exact content of the objective (in the example, essays)?

How To Establish Content Validity

Here is the most straightforward way to establish the content validity of the four-question test for each objective once you have written the questions:

1. Find two other teachers who are familiar with the content of the objective. (If they have taught the objective themselves, so much the better.) Ask them to serve as judges.

2. Ask each judge to examine each question independently to see whether it matches:

   -- *the exact content of the objective.*
   -- *the intellectual process called for by the objective.*

   Ask each judge to record ratings on a sheet like this:

Objective:_____

Question Number	Content Match	Intellectual Process Match
1	Yes No	Yes No
2	Yes No	Yes No
3	Yes No	Yes No
4	Yes No	Yes No

3. Tally the judges' ratings. Any "no" vote makes that question not valid.

4. If any questions were judged not valid, write some more questions and have those new questions judged. Keep writing and having questions judged until you have four questions that receive all "yes" votes by both judges.

5. Ask each judge to look independently at the four questions as a group and determine whether the four questions taken together represent the full range of learning called for by the objective. Both judges must agree that they do.

6. If the judges do not agree that the learning is well represented, write some new questions on whatever gaps the judges can help you identify, have those questions judged on content and intellectual process match, and have the improved group of four questions judged for representativeness. Keep writing and judging until a group of four questions is approved as representative by both judges.

You now have a four-question test that has content validity for an objective. This is a high standard to set: two independent judges in complete agreement. Validity of your test is that important if students are to be judged fairly.

Establishing Reliability of Paper-and-Pencil Testing

What Is Reliability?

Reliability is the extent to which a test can be depended upon to provide consistent, nearly error-free information. Reliability is expressed in degrees because the reliability number represents the likelihood that the students' score would be repeated if the assessment were readministered. Reliability can be considered the consistency among multiple administrations of the same assessment or the consistency among multiple raters in obtaining the same rating. For example, if students were tested on a Tuesday and again with the same test on Wednesday, and if there had been no further instruction or any other intervening factors such as a change in testing conditions, they should perform similarly on both days. Further, if three different teachers rated one single student performance according to a defined rating scale, their ratings should agree. The coefficient of 1.0 would represent perfect agreement and freedom from all errors between performances the first and second times. Conversely, 0.0 reliability indicates zero degree of agreement. A 1.0 is never reached in educational practice. For this discussion, we are roughly equating reliability and consistency, and thus illustrating a very informal determination of reliability. More formal procedures are detailed in the second chapter, "Ensuring Reliability" and in the sources cited in the annotated bibliography.

You ended the last section with a content-valid four-question test for an objective. Now it is time to calculate the reliability of students' test scores on that four-question test.

Preparing To Determine Consistency

First, you will need more test questions for each objective. Let us suggest that you develop an "equivalent" form of your four-question test--that is, four more questions written to match each objective (of course, those questions must also be judged to have content validity, just as the first four questions were).

You now have eight questions for each objective--a good pool of questions. Such a pool has many advantages, apart from its use in calculating reliability:

-- *You can use one question in a pretest to give you an idea about how much students know before you teach an objective.*

- *You can use one question in a daily quiz or a weekly review to give you an idea about how students are progressing while you are teaching an objective.*

- *You can give one question to students themselves to help them understand what it is they are supposed to be learning.*

- *You can give one question to parents to help them understand what their children are supposed to be learning.*

And you would still have four fresh content-valid questions left for the unit test or for the final exam.

Once you establish consistency using the eight questions with one class of your students, you will not need to do it again, unless your students change significantly (e.g., you get a much brighter class, you get a much slower class). If students are different enough, the consistency of their test scores could be different, thus forcing you to recalculate consistency.

Because you are interested in whether students are achieving the objectives, the reliability question is: How consistent are your decisions about students' achievement or nonachievement of each objective? Thus, you must calculate the percentage of consistent decisions for the two equivalent forms of the four-question tests for each objective.

How To Calculate Consistency

Here is a simple way to calculate informally consistency for the questions for each separate objective:

1. Divide the eight questions into two four-question tests. Let's call the two tests Form X and Form Y.

2. Let's establish the "achievement" standard as 75 percent--that is, students who get three or four questions correct on a four-question test are classified as "achievers." Students who get two or fewer questions correct on a four-question test are classified as "non-achievers." Here are the possibilities:

    a. Students are classified as "achievers" on both Form X and Form Y, getting three or four out of four questions correct on each form.

    b. Students are classified as "nonachievers" on both Form X and Form Y, getting two or fewer out of four questions correct on each form.

    c. Students are classified as "achievers" on one form and "nonachievers" on the other form, getting three or four questions out of four questions correct on one form, but not on the other.

3. Fill in the following boxes:

		Form Y	
		Achievers	Nonachievers
Form X	Achievers		
	Nonachievers		

Let's take an example. Say you have a class of 20 students. You classify 9 of them as achievers on both Form X and Form Y; 7 as nonachievers on both Form X and Form Y; 0 as achievers on Form X, but nonachievers on Form Y; and 4 as nonachievers on Form X, but achievers on Form Y. Here are the filled-in boxes:

		Form Y: Achievers	Form Y: Nonachievers
Form X	Achievers	9	0
	Nonachievers	4	7

9+0+4+7=20 total students

4. Fill in this formula, using the numbers in your completed boxes:

$$\frac{\text{Achievers on both forms} + \text{Nonachievers on both forms}}{\text{Total number of students}} \times 100 = \text{Percent consistency of decisions}$$

Let's use the example above:

$$\frac{9+7}{20} \times 100 = \frac{16}{20} \times 100 = \frac{1600}{20} = 80 \text{ percent consistency of decisions}$$

Thus, in our example, the decisions made about whether students were classified as achievers or nonachievers were consistent--or reliable--80 percent of the time.

How Much Consistency Is Enough?

Unfortunately, there are no guidelines established by statisticians or psychometricians for judging the percentages. So, you must resort to common sense.

Here are some factors you may wish to consider:

-- Your average standard for achievement. If 75% achievement level is considered average for your students, you may wish to adopt a 75% standard for consistency.

-- Your numerical standard for passing. If 60% achievement level is considered passing for your students, you may wish to adopt a 60% standard for consistency.

-- Consistency in the aggregate. Rather than requiring that each test has a consistency rating at a prescribed level, you may wish to take the average consistency rating of all tests that you are using, and establish a minimum.

For example, you might wish to say all tests will have a consistency rating of at least 60% and an aggregate average rating of 75%.

-- The extent of reliance upon locally written and designed criterion referenced tests. As you design the total learning assessment plan, locally developed tests must be supplemented with norm-referenced and standardized assessments. On any given objective, if you are relying solely on local instruments, the consistency rating must necessarily be higher.

For further discussion of this topic see pp. 36-37.

ESTABLISHING GOOD PERFORMANCE OR PRODUCT ASSESSMENTS

While many things can be assessed by paper-and-pencil tests, others cannot. Here are some that cannot:

Performances	Products (Artifacts)
playing a tuba	*a poem*
dancing	*a bookcase*
acting	*an omelette*
diving	*a watercolor*
speaking	*an essay*
swinging a bat	*a weld*

Notice that the ones in the first column are all performances. They have to be assessed "live" while they are still in progress. Once they are over, they are over.

Notice that the ones in the second column are all products (artifacts). They can be assessed after they have been produced. In fact, because they will last, they can be assessed long afterward--except for the omelette.

Of course, thanks to modern technology, you can record a performance--on audiotape or on videotape, for example. But recording a performance does not make it a product. It is still a performance, and you will have to assess it as a performance.

Assessing the Performance or the Product--or Both

Of course, all products result from performances. That means you have to decide whether to assess the performance, the product, or both. For example:

> *You can assess a student's technique for varnishing bookcase (her performance while she varnishes), or you can assess the varnish on the bookcase after it has dried (her product), or you can do both.*

The same thing applies to welding technique and to the finished weld, to poetry-writing technique and to the finished poem, to watercoloring-technique and to the finished watercolor.

Here are some advantages of assessing the performance rather than the product:

You can time the performance.
You can assess each separate part of the performance.
You can pinpoint correct techniques.
You can pinpoint incorrect techniques.

You can give immediate ratings.
The student is more likely to improve techniques because you have pinpointed good and bad ones at the time of such activity.

Here are some advantages of assessing the product rather than the performance:

The student can do her best work without the anxiety of being watched and the pressure of being timed.
You can assess each separate part of the product.
You can pinpoint weak elements in the product.
You can pinpoint strong elements in the product.
You can give careful ratings without the pressure of time.
You can use several judges more conveniently.
You can take more time to reflect yourself on what you observed and had to assess.

Recording the Performance Before Assessing It

Here are some advantages of recording a performance before assessing it:

You can give careful ratings without the pressure of time.
You can give several ratings over time.
You can use several judges more conveniently.

Here are some disadvantages of recording a performance before assessing it:

Recording it may affect the student's performance.
The recording will lose some of the warmth and intensity of the actual performance.
Neither the camera nor the microphone will have the flexibility of a skilled observer's eyes and ears. An observer can walk around the performer, for example, but the camera cannot.

PLEASE NOTE: You could assess the performance live and judge the tape later.

Building Archives of Performances and Products

We advise you to retain tapes of performances, or copies, or photographs of products so that you can anchor your rating from year to year. In a subsequent year as the assessment nears, view the performance tapes and photographs from the past as a rating retooling. You will not begin "cold" with the first performance that way.

Measuring Performances

Few aspects of any performance can be measured. Most have to be judged. But here are a few examples of measuring performances:

Keeping score at a game.
Using a stopwatch to time a sprint.
Using an audiometer to measure applause.
Using an emissions testing machine to check a tune-up.
Counting the bullet holes in each concentric ring of a bull's eye.
Using calipers to test the diameter of a rod.

Unfortunately, although measurement is inexpensive, faster, more accurate, less subject to argument between teacher and student, more reliable, and more valid than judgment, performance assessment usually depends on judgment. So measure when you can; judge when you can't.

Considerations about Judging

Here are some factors to keep in mind when using judgment:

No two judges will necessarily look for the same parts of the performance or the product.
No two judges will necessarily have the same opinion about any part of the performance or the product.
No two judges will necessarily give the same weight to any part of the performance or the product when reaching an overall judgment.
No one judge will necessarily look at the same parts or have the same opinion or give the same weights several days apart.
Judges find it difficult to separate the performance or the product from the student who produces it if they know anything about the student. They may rate the student as well as the performance or product.

One way to reduce these variations is to use rating scales.

Establishing Good Rating Scales

Rating scales must be used to increase both the validity and the reliability of judgments. A scale is a line with points arranged in a sequence which can be from small to large, cold to hot, empty to full, ugly to beautiful, awkward to graceful, light to heavy, smooth to rough, quiet to noisy. The points must be in a sequence for a scale.

A measuring scale is for things that can be measured. (We mentioned that earlier.) We have good measuring devices with good scales for length, for weight, for temperature, for time, for blood pressure, and for angles--among other things. Good measuring devices with good scales include meter sticks, bathroom "scales," thermometers, clocks, sphygmomanometers, and protractors--among others.

A rating scale is for things that cannot be measured but must be judged. To judge something is to slide it along a scale and stop at the point where it matches something on the scale. The stopping point is the rating.

Now, that may sound extreme. But just think about it. Every judgment is a comparison. You cannot compare without some sort of scale. It may be in writing, or it may be in your mind. It may have a few numbered points, or it may have a thousand unnumbered points.

Which of two performances or products is tastier, better composed, more graceful, better proportioned, more pleasing to the audience, better joined, more beautiful? You cannot answer without comparing the two on identical elements. Further, if you were to compare more than two performances or products, again you would use an identical element as a comparative scale. It would be meaningless to say one product is beautiful and the other is tasty.

Every one of the following statements is a judgment. That is, the person making the statement had to slide something along a scale and stop at the point where it matched:

That is the most beautiful sunset I have ever seen.
He can adjust a carburetor about as well as the usual kid.
He lacks grace.
Even allowing for the fact that he has been playing the trombone only three years, I would say he is hopeless.
I'll give the watercolor an A+.
Looks like varsity material to me.
Talk about pound cake!
Good on the approach, excellent on the takeoff, blew it on the entry.

Admittedly, it is not easy to slide a sunset or even a football player back and forth on a scale until it matches some point. You have to do it in your mind. But you have to do it or you cannot judge.

Before you can create a rating scale yourself, you have to make three decisions:

1. What will the scale ask the teacher to judge?
2. How many points will the scale have?
3. How will the points be labelled?

What Will the Scale Ask the Teacher To Judge?

The name of the scale tells the teacher what trait to look for: symmetry, flavor, backswing, texture, grip, tone, smoke, splits, consistency, presence, safety steps, resonance, persuasiveness, posture--whatever is worth rating because it is an important part of the performance or the product. The name of the scale makes the teacher look for the same trait every time. (A more complete discussion of traits can be found in the third chapter, "Performance Assessment.")

How Many Points Will the Scale Have?

Ultimately criterion-referenced assessment has two points on the scale. Either the performance or product is acceptable and demonstrates the student has mastered the objective or the performance or product is unacceptable and demonstrates the student failed to master the objective. Whether you choose to use an even numbered scale of 100 points or one with just two points, you need to have a cutoff point. For example, if, theoretically, you had a one hundred point scale, you might establish a cutoff point as 60 (or passing). Teachers are put in a "forced choice" position. They must decide on an even point scale whether the performance or product has sufficient merit that the student has achieved the objective.

The more points you use, the more difficult it is for the teacher. You should not use more than about eight unless you really need more precise distinctions--and unless the teacher is an expert and can really make such fine distinctions accurately. Remember, you can use more than one teacher to assess a performance or product and add scores together for a cumulative scale. Here at Homewood-Flossmoor High School, for example, we use a 1 through 4 rating scale for our writing assessment; two raters read each paper; thus our cumulative scale is 2 through 8.

You can use an odd number of points to make it easy for the teacher to give the midpoint to the average performance or product; you can use an even number of

points to force the teacher to decide whether the performance or product is above or below average.

How Will the Points Be Labelled?

Labelling has three possibilities. You probably cannot use the best one because of time, cost, and convenience; so you will have to trade down to one you can afford.

Actual performances or products. This is the best one. All the others are substitutes. Imagine a scale, say a six-point scale six feet long. Imagine it nailed to the classroom wall with six wooden pegs one foot apart. Hanging from each peg is an actual product or a tape recording of an actual performance, with the worst on one end and the best on the other. The products could be essays, seams, or ceramics; the performance recordings could be clarinet solos, dances, or dives. A teacher would use the scale by examining student's product or watching a student's performance and then matching it to the closest one on the scale. The match would determine the rating. The performances or products become anchor points for the teacher.

Every teacher must have such a scale of actual performances or products in her mind--if not on the wall.

Descriptions of performances or products. Such a scale simply substitutes written descriptions for actual performances or products at each point on the scale. This scale gives teachers less guidance than actual performances or products, but still gives them an adequate amount. The descriptions might sound like this for an arabesque:

1 - Extended leg's knee bent and toe not pointed; supporting leg wobbly; arms floppy; torso slumped.

2 - Extended leg's knee straight, but toe not pointed; supporting leg wobbly; arms and torso lifted.

3 - Extended leg's knee straight and toe pointed; supporting leg solid; arms strong; back arched; head lifted.

4 - Extended leg's knee straight and toe pointed, with foot higher than head; supporting leg solid; arms expressive; back arched; head tilted expressively.

Numbers. Such a scale simply substitutes numbers for the written descriptions of actual performances or products at each point on the scale. (This scale gives teachers the least amount of guidance.) In using such a scale, every teacher must imagine her own scale of actual performances or products--or written description of those performances or products--and equate it to the numbers. The danger, of course, is that each teacher will have a different imagined scale and equate it to the numbers in a different way.

As you move from actual performances or products to descriptions of those performances or products to numbers, the teacher's ratings become less valid and in all likelihood less reliable. Using just numbers to mark the points on your scales will make it too difficult to prove the validity and reliability of your performance or product assessments. Further, it will be very difficult to explain these ratings to students and parents.

How Many Good Rating Scales?

This is a matter of judgment. It may be that you can combine the components of a performance or product into a single rating scale--as in our arabesque example earlier, where the teacher gives one rating to the arabesque, taking into consideration legs, arm, torso, and head all as one.

However, especially for less experienced teachers, it is better to use a separate rating scale for each component of the performance or product. In the case of our arabesque, teachers would have one scale for the position of the extended leg, one for the strength of the supporting leg, one for the grace of the arms, one for the position of the torso, and one for the use of the head. This practice is fairly common in judging essays, where the writing might be rated on one scale for organization, on another for creativity, on another for grammatical correctness, and so on.

Even in the Olympics--where there are expert judges--skaters get ratings on two components: one for technical performance and one for style. Of course, their colleagues on the diving boards and on the parallel bars get just one rating, combining everything about the dive or the routine.

So, the answer to how many rating scales to use is not clear. Several scales are probably more accurate than one, provide better diagnostic information to a teacher, and are likely to reflect the totality of instruction; use several for each performance or product. Further, because performance appraisal is very time consuming, applying several scales to one performance is an efficient answer as well.

Establishing Validity of Performance or Product Assessments

As with paper-and-pencil test questions, you must establish the validity of the rating scales used to judge a performance or product. To say that the rating scales have content validity is to say that they adequately represent the learning called for by the objective that specifies a given performance or product.

You can already see how using several rating scales to judge a performance or product--rather than just one overall scale--helps you justify your "test" as adequately representing the learning called for. If the objective were "Does an arabesque," then five scales would ensure that you were judging five different components (extended leg, supporting leg, arms, torso, and head)--a good representation of the overall arabesque. Note: Now is not the time to forge new vocabulary or content in your discipline. You should be identifying traits which are commonly agreed upon standards of quality performance.

How To Establish Content Validity

Here is the most straightforward way to establish the content validity of several rating scales for judging a performance or product called for by an objective:

1. Find two other teachers who are familiar with the content of the objective--that is, the performance or product. (If they have taught the objective themselves, so much the better.) Ask them to serve as judges.

2. Ask each teacher to examine each rating scale independently to see whether it rates an important component of the performance or product:

Ask each teacher to record ratings on a sheet like this:

Objective: _____

Name of Rating Scale	Importance of the Component
_____	High Low
_____	High Low
_____	High Low
_____	High Low

3. Tally the teachers' ratings. Any "low" vote makes that rating scale invalid.

4. If any rating scale were judged of low importance, omit it.

5. Ask each teacher to look independently at the remaining rating scales as a group and determine whether the rating scales taken together represent the full range of learning called for by the objective. Both teachers must agree that they do.

6. If the teachers do not agree that the learning is well represented, write some new rating scales for whatever gaps the teachers can help you identify, have those rating scales judged on importance, and have the improved group of rating scales judged for representativeness. Keep writing and judging until a group of rating scales is approved as representative by both teachers.

You now have a group of rating scales that has content validity for an objective. This is a high standard to set: two independent judges in complete agreement. But remember: Validity of your test is that important if students are to be judged fairly.

Establishing Reliability of Performance or Product Assessments

In order for the performance or product ratings to be reliable, the teachers themselves must possess certain characteristics:

Subject matter knowledge --Teachers must be extremely knowledgeable about the performances or products being judged, and the scales being applied. They should have a clear understanding of a "perfect" performance or product, and a clear definition of achievement of the objective.

Technical expertise in using rating scales --Teachers must be free from the most common errors people make when using rating scales. These are described in more detail on pages 44-45 of this volume.

In addition to these characteristics, the teachers must be "of one mind" in their ratings of the performances or products--that is, all teachers should give the performances or products the same, or nearly the same, ratings. The consistency of the ratings is what reliability is all about. To ensure that consistency, training is usually required.

Training the Teachers to Use Rating Scales

Training teachers to be consistent in their ratings among themselves should include these common sense activities:

Discussions of the rating scales being used so that interpretations of those scales are as similar as possible. (You can see how much easier this would be if the scales included actual performances or products or written descriptions of those performances or products.)

Opportunities to judge performances (by using videotapes, in all likelihood) and to discuss the ratings each teacher gave.

Discussions to iron out significant differences in the ratings the teachers gave--that is, differences of two points or more (if you're using a scale of four through eight points).

Identification and elimination of teachers who, by the end of the training, do not have the three characteristics listed previously: subject matter knowledge, technical expertise in using rating scales, and personal objectivity.

Not surprisingly, teachers who once rated performances or products consistently among themselves might over time become inconsistent with each other. Thus, annual brush-up training may be necessary. For more information, see pages 45-46.

Calculating Reliability

Two different cases must be considered: (1) ratings when only one teacher is available; and (2) ratings when two or more teachers are available.

Whenever possible, two or more teachers should be used. The advantage is obvious: A student is inevitably judged more fairly if more than one person is doing the judging. The disadvantage is equally obvious: Sometimes, more than one qualified teacher is simply not available--for example, a high school may have only one instrumental music teacher who knows about strings. Getting another strings teacher from another school district or a concert violinist from town is rarely feasible. Therefore, we know that there will be times when one teacher must suffice.

Reliability of Ratings by One Teacher

You could simlulate two teachers rating a performance by videotaping the performance and having your one available teacher rate it again later. However, videotaping equipment is not available in every school district, either. So, if you have

only one qualified teacher and videotaping is unavailable, use this procedure to establish the reliability of the single teacher's ratings:

1. Choose another teacher (of any subject) or an administrator to serve as a "monitor" of the judging process.

2. Have the one qualified teacher (say, your only strings teacher) use the rating scales (say, for strings solos) in the presence of the monitor.

3. When the performance is finished, have the teacher explain the reason for each of her ratings to the monitor.

4. Have the monitor check to be sure that the teacher displayed sufficient subject matter knowledge, technical expertise in using the rating scales, and personal objectivity in judging the performance.

5. Have the monitor vouch for the reliability of the ratings given by the single teacher.

You can easily see that this procedure is not "calculating" reliability at all, but rather "judging" it. Nonetheless, it is the only practical alternative when you have only one qualified teacher and you must rate performances.

When videotaping equipment for performances is available or when the thing to be rated is a product--rather than a performance--and therefore lasting, you could use the procedure in the following section to calculate reliability (by using two different ratings by the same teacher, but separated by a few days, and acting as if the two ratings were by two different teachers). However, it seems more straightforward even in these cases just to follow the procedure above.

Reliability of Ratings by Two or More Teachers

Let's start with two judges. Here is the procedure to use to calculate reliability:

1. List the ratings given to each student by each teacher in a table like this:

Student Number	Ratings by:	
	Teacher A	Teacher B
1	6	5
2	5	5
3	3	4
4	4	3
5	7	6
6	3	6
7	3	3
8	2	3
9	1	2
10	5	7

2. Add another column to show the difference (in number of points) between the ratings of Teacher A and Teacher B:

Student Number	Ratings by:		Difference
	Teacher A	Teacher B	
1	6	5	1
2	5	5	0
3	3	4	1
4	4	3	1
5	7	6	1
6	3	6	3
7	3	3	0
8	2	3	1
9	1	2	1
10	5	7	2

3. Fill in the following formula to calculate the percent of consistent ratings between Teacher A and Teacher B:

$$100 \times \frac{\text{Number of differences less than 2}}{\text{Number of students rated}} = \text{Percent of consistent ratings}$$

$$100 \times \frac{8}{10} = \frac{800}{10} = 80$$

percent consistent ratings between Teacher A and Teacher B

When there are more than two teachers, simply determine the consistency of ratings for two teachers at a time, using the procedure above. For example, if there are three teachers (Teacher A, Teacher B, and Teacher C):

1. Determine the consistency of ratings between Teacher A and Teacher B.
2. Determine the consistency of ratings between Teacher B and Teacher C.
3. Determine the consistency of ratings between Teacher A and Teacher C.
4. All three calculations should exceed your acceptable reliability score.

Unfortunately, statisticians and psychometricians have established no firm guidelines for judging how consistent the ratings should be to be determined "reliable." The choice is one you will have to make. We recommend that it be a collective decision with other people in your district.

ENSURING RELIABILITY

ESTABLISHING RELIABILITY FOR PAPER AND PENCIL TESTING

We cannot single out a method and say it is only applicable for performance or for paper and pencil testing. We do, however, have some peculiarities in paper and pencil tests which we may not necessarily have in performance ratings. There are several distinct kinds of test reliability:

Test-Retest--Administer the same test to the same students on two different days.

Parallel Forms--Administer two forms of the same test to the same students (either on the same day or on two different days)

Split Half--Subdividing a single test into halves (usually the odd-numbered questions versus the even-numbered questions) after the test has been administered in one sitting

Internal Consistency--Indexes of reliability are based upon the consistency among the questions on the test.

Theoretical Considerations

In writing this guide, we first ruled out test-retest and parallel forms as the more impractical for teacher-made and district-made tests. We considered each of these methods too intrusive to an instructional program and unwieldly in an individual classroom K-12. These methods are more appropriate for standardized tests. We did, however, rule them out reluctantly. They probably would have been more accurate than the latter two. (Refer to Gronlund or one of the other standard sources in the annotated bibliography for good descriptive methods.) The two we kept, split-half and internal consistency may be thought of as compromises because they're practical and because they can be considered as estimates of the first two. This is clear in the split-half case. The half-tests operate as parallel forms of the test given on the same day. Under any circumstances, reliability indexes for paper and pencil tests at least give (albeit roughly) some indication of whether we are on target in making accurate assessment.

Ironically, the four reliability methods above are considered "classic" and were developed a long time ago for the long-used norm-referenced tests. They are based upon two assumptions in norm-referenced tests: 1) the test scores have a relatively wide range (the same as saying that the scores are widely distributed along a scale)

and 2) the variability of the scores is relatively large. It follows from these assumptions that the mean score must be somewhere in the middle of the range of possible scores. By contrast, criterion-referenced testing typically has restricted range, small variability, and means at the high end of the scale. We want our students to meet the criteria. Psychometric scholars think that the classic methods are not very applicable to criterion-referenced testing. Perhaps the best rule for us to follow is to decide whether we can reasonably meet the assumptions, and then go ahead and use the classic methods for criterion-referenced testing. There is (good news) an alternative to the classic.

If we think of criterion-referenced tests as "mastery" tests, some authentically simple methods count the number of "masters" and "non-masters" on two forms of a test. Recall that we may never have more than one form of a locally-made test, but if we use split halves in place of alternate forms, we can use some simple indexes of various percentages. Some of these are illustrated in the Exhibits.

Four reliability or consistency methods are presented in the Exhibits.

Consistency Methods

1) Exhibit B--shows step-by-step how to compute **internal-consistency reliability** for 20 students on a 5-question objective test (scored "1" for right and "0" for wrong on each question).

2) Exhibit D--shows the computing steps for **split-half reliability** for 20 students on a 20 question test, split into two halves of 10 questions each. Scores shown are the number of questions (out of 10 questions on a form) answered correctly.

3) Exhibits E and F--show the computations for two short methods for assessing the consistency with which two halves of a test classify students as "masters" or "non-masters." We call these **split-halves mastery consistency methods.**

Internal Consistency Reliability--Exhibit B

Step 1

Exhibit B begins with a Data Box in which the entries are "1's" (ones) for particular students passing particular questions and "0's" (zeros) where particular students failed particular questions. In 76 instances some question was passed by some student. 76 is our Check Sum for row totals and column totals. Add ones down each column to obtain the column totals (13, 13, 16, 15 and 19.) Add these up and obtain 76. Add across the columns for each student and obtain in totals 1, 2, 3, or 5. Add these up and again obtain 76, showing that we have done our addition correctly. Square each Question Total (e.g., $13^2 = 169$) and add up all five squares, obtaining

1,180. Square each of the student totals and record them in the (Student Total)^2 column. Add them up and obtain 350. Square 76, the sum of 76 scores, and obtain 5,776.

EXHIBIT B--Computing Diagram for Internal Consistency Reliability

STEP 1.

Student No.	QUESTION NO. A	B	C	D	E	Student Totals	Squared Student Totals
1	1	1	1	1	1	5	25
2	1	1	1	1	1	5	25
3	1	1	1	1	1	5	25
4	1	1	1	1	1	5	25
5	1	1	1	1	1	5	25
6	0	0	0	0	1	1	1
7	1	1	1	1	1	5	25
8	1	1	1	1	1	5	25
9	1	1	1	1	1	5	25
10	1	1	1	1	1	5	25
11	1	1	1	1	1	5	25
12	1	1	1	1	1	5	25
13	0	0	1	1	1	3	9
14	0	0	1	1	1	3	9
15	0	0	0	0	1	1	1
16	0	0	0	0	1	1	1
17	0	0	0	0	0	0	0
18	1	1	1	1	1	5	25
19	1	1	1	1	1	5	25
20	0	0	1	0	1	2	4
QUESTION TOTALS	13	13	16	15	19	76	350
SQUARED TOTAL	169	169	256	225	361	5776	

Sum of (Question Totals Squared) = 1,180

Step 2

Now we have all the data necessary to fill in the appropriate formula, Reliability Equation:

$$\text{Reliability} = 1 - \frac{\frac{D}{(n-1)(N-1)}}{\frac{A}{N-1}}$$

Compute the four terms A, B, C, D for the Reliability Equation. Divide the Sum of the Squares of the Student Totals (350) by the number of questions (5), giving 70.

Divide the square of 76 (5,776) by the product of number of questions and number of students (5 x 20 = 100), yielding 57.76. Subtract 57.76 from 70, giving 12.24 for Term A.

To compute Term B divide the Sum of the Square of the Question Totals (1,180) by the number of students (20) getting 59. Subtract from 59 the same second value from Term A (57.76) getting 1.24 for Term B.

To compute Term C, start first within the brackets. Multiply the number of questions by the number of students, 5 x 20, and obtain 100. Subtract from 100 the Sum of Student Totals (76), getting 24. Multiply 24 x 76, yielding 1,824. Divide 1,824 by 100 (5 x 20) and get 18.24 for Term C.

Term D is obtained by subtracting from Term C both Term A and Term B. Thus, 18.24 - 12.24 - 1.24 - 4.76, Term D.

Continuing with Step 2 of Exhibit B, we are ready to substitute terms in the Reliability Equation. The equation has two terms, "1" and "Something" (which looks complicated). Let's compute the "Something." We use only Terms D and A from above plus new terms, "n" (the number of questions) and "N" (the number of students). Thus n = 5 and N = 20. First, compute n - 1 and N - 1; they are 4 and 19. Divide Term D by the product (4) (19). This gives the numerator 4.76/76 = .0626. The denominator is Term A divided by N - 1, 12.24 divided by 19 which reduces to .6442. Dividing the numerator (.0626) by the denominator (.6442) give us .0972, which is the "Something" we needed. Subtracting .0972 from 1 gives us the Internal Consistency Reliability = .90.

EXHIBIT B (cont)

STEP 2 Terms for Reliability Equation

$A = 1/5(350) - 76^2/(5)(20) = 70 - 5{,}776/100 = 70 - 57.76 = 12.24$
$B = 1/20(1{,}180) - 76^2/(5)(20) = 59 - 5{,}776/100 = 59 - 57.76 = 1.24$
$C = (76)[(5)(20) - 76]/(5)(20) = (76)(24)/100 = 1{,}824/100 = 18.24$
$D = C - A - B = 18.24 - 12.24 - 1.24 = 4.76$

$$\text{Reliability} = 1 - \frac{\frac{D}{(n-1)(N-1)}}{\frac{A}{N-1}} = 1 - \frac{\frac{4.76}{(4)(19)}}{\frac{12.24}{19}} = 1 - \frac{\frac{4.76}{76}}{\frac{12.24}{19}} = 1 - \frac{.0626}{.6442} = 1 - .0972$$

Reliability = .90

Split-Half Reliability for a Test ("Long" Method)--Exhibit D

Step 1

Exhibit D begins with a Data Box in which the entries are odd half scores and even half scores for students. Each student took an apparent whole test, but in the

scoring process later the test was scored by halves. Thus each student has an odd score and also an even score. Having entered the pair of scores for each student, we next take the difference between the pair of scores for each student. Consistently we subtract the Even score from the Odd score to get the difference (Diff). In the next column we square each of these differences. In the next column we add the odd and even scores in each pair for each student. In the far right column we square the sums of the scores for each student. Now, we add up all the columns. Please note: there are no row totals in this method. We do, however, have two methods to check our arithmethic. First, the sum of the difference ought to equal the actual difference between the odd and even halves. In our example, the two column totals for the odd and even scores (117 and 136) must yield a difference of -19. Also, the sum of the odd and even halves ought to equal the Odd-Even student totals. The two column totals for the odd and even scores (117 and 136) must yield a sum of 253.

EXHIBIT D--Split-Half Reliability for a Test

STEP 1 Totals for Scores and Squares

Student No.	Halves of Test Odd	Even	Diff	$Diff^2$	Odd-Even Student Totals	Student Total Sq.
1	6	7	-1	1	13	169
2	6	8	-2	4	14	196
3	7	7	0	0	14	196
4	7	9	-2	4	16	256
5	7	7	0	0	14	196
6	3	4	-1	1	7	49
7	6	7	-1	1	13	169
8	9	10	-1	1	19	361
9	8	9	-1	1	17	289
10	7	9	-2	4	16	256
11	8	9	-1	1	17	289
12	7	9	-2	4	16	256
13	4	5	-1	1	9	81
14	4	5	-1	1	9	81
15	3	4	-1	1	7	49
16	3	5	-2	4	8	64
17	3	3	0	0	6	36
18	7	7	0	0	14	196
19	7	7	0	0	14	196
20	5	5	0	0	10	100
Totals	117	136	-19	29	253	3485

Step 2

Compute the two terms A and B. Divide the Sum of the $(Diff)^2$ column (29) by the number of students (20) and get 1.45. Divide -19 by 20, getting -.95. Square -.95, getting .9025. Subtract .9025 from 1.45 getting .5475, which is Term A. To find Term B first divide 3,485 (Sum of Squares of Student Totals) by 20 (number of students) to get 174.25. Divide Sum of Odd + Even scores (Student Totals), 253, by 20 (number of

students) giving us 12.65, which when squared gives us 160.0225. Subtract 160.0225 from 174.25 to obtain 14.2275, which is Term B. Find A/B by dividing .5475 by 14.2275 to get .0384. Finally, subtract .0384 from 1 to obtain .96, the Split-Half Reliability ("Long" Method).

EXHIBIT D (cont)

STEP 2 Terms for Reliability Equation

$A = 29/20 - (-19/20)^2 = 1.45 - .9025 = 5.475$
$B = 3,485/20 - (253/20)^2 = 174.25 - 160.0225 = 14.2275$

Reliability = 1 - A/B = 1 - .5475/14.2275 = 1 - .0384 = .96

Please Note: This method assumes that the "odd" and "even" halves are equivalent. In some tests item arrangement may preclude using this method.

Reliability Index

Exhibit G--shows a set of guidelines for the expected size of reliability indexes for paper-and-pencil locally-made tests of varying lengths.

EXHIBIT G--Expected Size of Reliability Indexes for Carefully-Made, Locally-Made Tests of Various Lengths
(Assuming at least moderate score range and moderate score variability)

Length of Test in No. of Questions	Expected Size of Reliability Indexes: Average Size	Range in Size
More than 30	Mid .90's	.90's
30	.87	Mid .80's to High 80's
20	.80	Mid .70's to Mid 80's
15	.75	High 60's to Low 80's
5	.55	Mid 40's to Mid .60's

As a guideline for interpreting obtained reliability coefficients from tests and rating scales, Exhibit G presents a brief table of the expected sizes of internal consistency test reliability. The coefficients covered by the table include Exhibits A, B, and D and Case 1, Exhibit C. The table in Exhibit G is set up with selected test lengths (in terms of number of questions) as one of the parameters. This was based upon the well-known fact that--other things being equal--the longer a test, the higher the reliability and vice versa. Therefore, the length factor of a test must be considered when interpreting a given reliability coefficient. Notice our assumptions that a test is carefully made by local school personnel and that the scores yield a distribution with at least moderate score range and moderate score variability.

We developed the table on the basis of many observed reliabilities from our collective experience of tests which were "carefully-made," although we were not able to factor in observed reliabilities from competency based (or "criterion-referenced" or "mastery tests").

The table in Exhibit G refers only to reliabilities for tests; it does **not** include reliabilities for performance ratings. However, our experience shows that our performance rating reliabilities have been running in the .80's and .90's.

Therefore, Exhibit G should be considered as a rough guide to the interpretation of test reliabilities. Also, the table is tentative. In a year or two, when many schools have had the opportunities to use our handbooks and have fed back reliability data to us, we will be in a position to consider whether to revise the table.

By coincidence, as we were going to press, a relevant paper appeared in a measurement journal (Frisbie, 1988--listed in the Chapter Bibliography, p. 55). Frisbie has provided material suitable for a short course or workshop to teach classroom teachers to assess the reliability of their own tests. His context differs from ours in that (1) he is oriented toward tests for course grading; and (2) he addresses teachers in general.

One of the bottom lines for Frisbie is the rather strong statement he makes that the reliabilities of the teachers' own tests average about .50. This is considerably lower than we expect in our different context.

Split-Halves: Mastery Consistency Methods--Exhibits D and E

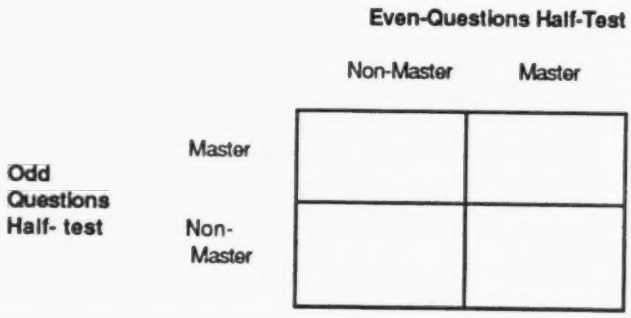
We now come to the previously mentioned short methods for determining the consistency with which two halves of a test classify students as "masters" or as "non-masters." We cannot use the scores on the half-tests exactly as they appear in Exhibit D, but we preserve some of the information in the Exhibit D data by a process of transformation of data. Having established a standard for mastery (we used 70% here), we can classify (on either half-test) the number of persons who reached the standard and the number who did not reach the standard. Our scores in Exhibit D are all one-digit numbers, the maximum possible score is 10 and 70% of 10 is 7. Therefore, we use the following code to classify the data from Exhibit D as "Master" or "non-Master." Masters earn "Odd" or "Even" scores of 7, 8, 9, 10; Non-Masters earn "Odd" or "Even" scores of 6, 5, 4, 3.

EXHIBIT D--Split-Half Reliability for a Test

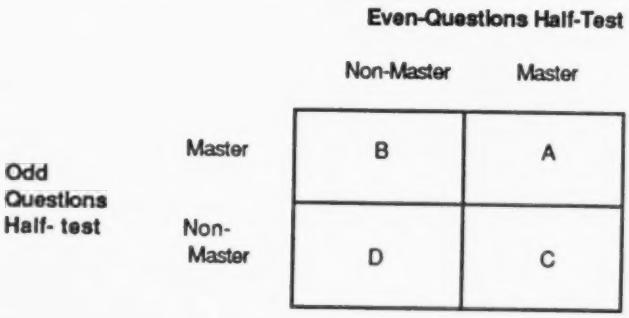
STEP 1 Totals for Scores and Squares

Student No.	Halves of Test Odd	Even	Diff	$Diff^2$	Odd-Even Student Totals	Student Total Sq.
1	6	7	-1	1	13	169
2	6	8	-2	4	14	196
3	7	7	0	0	14	196
4	7	9	-2	4	16	256
5	7	7	0	0	14	196
6	3	4	-1	1	7	49
7	6	7	-1	1	13	169
8	9	10	-1	1	19	361
9	8	9	-1	1	17	289
10	7	9	-2	4	16	256
11	8	9	-1	1	17	289
12	7	9	-2	4	16	256
13	4	5	-1	1	9	81
14	4	5	-1	1	9	81
15	3	4	-1	1	7	49
16	3	5	-2	4	8	64
17	3	3	0	0	6	36
18	7	7	0	0	14	196
19	7	7	0	0	14	196
20	5	5	0	0	10	100
Totals	117	136	-19	29	253	3485

Set up a new type of Data Box with two rows and two columns such as the one shown in Exhibit E. Label the top row as "Master" (on the Odd Test) and the second (or right) column as "Master" (on the Even Test). Label the bottom row as "Non-Master" (on the odd test) and the first (or left) column as "Non-Master" (on the even test).



Next, label the four cells of the 2 by 2 table as A, B, C, and D in the pattern shown in Exhibit E. Label the upper right cell as A (Master-Master). Label the lower left cell as D (Non-Master--Non-Master). Label the upper left cell as B (Master--Non-Master). Label the lower right cell as C (Non-Master--Master).



Now tally the instances of the four combinations in the four cells and then convert to Arabic numerals, yielding for our data: A = 9, D = 7, B = 0, and C = 4.

Compute the row and column totals for pairs of cell frequencies: A + B = 9; C + D = 11; A + C = 13; B + D = 7. Use the Check Sum of 20 (number of students) to verify that the sum of the pairs of cells for rows equals the sum of the pairs of cells for columns and finally that the sum of all four cells equals the Check Sum.

EXHIBIT E--Mastery Consistency for Split-Halves of a Test (Data of Exhibit D)

Code: Masters earn "Odd" or "Even" scores of 7,8,9,10;

Nonmasters earn "Odd" or "Even" Scores of 6,5,4,3.

STEP 1 Tally Performance Patterns on Two Halves

		Even-Questions Half-Test: Non-Master	Even-Questions Half-Test: Master	Totals
Odd Questions Half-test	Master	B = 0	A = 9	A+B = 9
	Non-Master	D = 7	C = 4	C+D = 11
Totals		B+D = 7	A+C = 13	A+B+C+D = 20

Consistency Indexes for Criterion-Referenced (Mastery) Tests

In the use of mastery tests, rather than estimates of student scores on a numerical scale of multiple points, our only purpose is to make a decision as to whether the student is to be classified as "master" or "non-master." Thus, new kinds of statistical indexes must now be introduced. The reliability indexes we saw previously are not sufficient unto themselves. However, a reliability estimate still provides partial information on the adequacy of a test in this respect; a very low value of reliability may disqualify a test for use, while a high reliability gives us some assurance that the test has some accuracy.

The indexes of consistency which we present here and recommend are these:

(1) Phi Coefficient
(2) Proportion of Consistency (from any and all sources), symbolized P_T, T for Total.
(3) Proportion of Consistency (beyond chance which is added by the test), symbolized by P_I, I for Instrument.

Phi Coefficient

With our data in hand from Step 1 in Exhibit E, we calculate the phi coefficient. We employ the phi coefficient (from statistics) here to summarize our mastery/non-mastery data. Verbally, the formula for phi (a Greek letter) is equal to the numerator (product of cells A and D minus the product of cells B and C) divided by the denominator (square root of the product of the four totals for rows and columns). Substituting, we have the numerator (9) (7) - (0) (4) = 63 and for the denominator the

square root of 9,009 (the product of 9, 11, 13, and 7), which equals 94.92. Dividing 63 by 94.92 yields the value of .66 for phi.

EXHIBIT E (cont)

STEP 2 Calculation of Phi Coefficient

$$\text{Phi} = \frac{AD-BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}} = \frac{(9)(7)-(0)(4)}{\sqrt{(9)(11)(13)(7)}}$$

$$\frac{63-0}{\sqrt{9{,}009}} = \frac{63}{94.92}$$

Phi = .66

A Phi coefficient of +.50 or more indicates that a test has adequate reliability for use.

P_T

Using the cell counts from Exhibit E, we substitute in the formula for P_T. We start inside the brackets first by subtracting from the Number of Masters on Both Halves, the Number of Non-Masters on Both Halves. Then we divide this difference by the Total Number of Students in Group. Substituting data, we have inside the bracket 9 + 7 = 16, which is divided by 20. This yields a value of .80, the proportion of decision consistency from all sources (in this case from chance and from the instrument). It is permissable to convert the P_T to a percentage. Thus, the interpretation here is that 80% of the students were consistently classified as "Masters" or "Non-Masters" by the two halves of the test.

EXHIBIT F--Proportion of Consistency Indexes (Data of Exhibit E)

STEP 1 P_T - Proportion of Consistency from all sources

$$P_T = \frac{\text{(Number of Masters) on Both Halves} + \text{(Number of Non-Masters) on Both Halves}}{\text{Total Number in Group}}$$

$P_T = [9+7/20] = [16/20] = .80$

No guidelines are available for interpreting the P_T index.

P_I

In order to calculate P_I (Kappa), it is necessary to first calculate an intermediate term P_C. This has been done for our data in Step 2 of Exhibit F. In Step 2 of Exhibit F the four row and column totals are substituted in numerators of the four proportions, where each proportion denominator is 20, the number of students. In the first of the two terms to be added we have the product of 9 and 13 divided by the square of 20, yielding 117/400. In the second term the product of 11 and 7 is divided by the square of 20, yielding 77/400. Combining the two terms with a common denominator, we have 194/400, yielding .48 as P_C.

EXHIBIT F (cont)

STEP 2 P_C - Proportion of Consistency from Chance Alone

P_C = (Proportion of Masters on Odd Half) *
(Proportion of Masters on Even Half) +
(Proportion of Non-Masters on Odd Half) *
(Porportion of Non-Masters on Even Half)

=

(9/20) (13/20) + (11/20) (7/20) =
(9) (13)/(20) (20) + (11) (7)/(20) (20) =

117/400 + 77/400 = 194/400 = .48

In Step 3 we substitute the two proportions just calculated in the simple formula for PI. In the numerator we subtract P_C from P_T, .80 - .48 = .32. In the denominator we subtract P_C from 1, yielding .52. Dividing .32 by .52 gives the value of .62 for P_I. It is permissable to convert P_I as a percentage. We may interpret the .62 as meaning that the test contributes 62% of the consistency of decisions.

EXHIBIT F (cont)

STEP 3 P_I-Proportion of Consistency Added by Test

NOTE: P_I is identical with "Cohen's Kappa Coefficient" from the literature of Measurement Statistics.

$$P_I = \frac{P_T - P_C}{1 - P_C} = \frac{.80 - .48}{1 - .48} = \frac{.32}{.52} = .62$$

CAUTION: P_T and P_I are not interchangeable; they do not assess the same thing. The choice of which to use depends upon whether you wish to focus on the total consistent classification or whether you are more concerned about the degree of consistency contributed by the measuring procedure itself.

ESTABLISHING RELIABILITY OF PERFORMANCE RATINGS

We need to set the stage for and to preview some of the more useful kinds of indexes for reliability and for agreement of raters for assessing the adequacy of performance rating scales. The choice itself of an appropriate reliability index (that is, a technique) for looking at the adequacy of raters and rating instruments depends upon two factors: (1) Performance Conditions; and (2) Number and Duration of Performances.

Performance Conditions

We consider here what is probably the most typical performance condition, that is, a rating scale with a concise verbal description of the performance to be rated (such as the quality of a basketball lay-up performance) and a run of consecutive whole digits to indicate points on a scale of quality. Typically, most scales have numbers such as 1,2,3; 1,2,3,4; 1,2,3,4,5; 1,2,3,4,5,6; or 1,2,3,4,5,6,7.

We would choose a scale with an odd number of numerical points if we want to have a middle (median) number which is an "undecided" or "average" or something else "on the fence." We would choose a scale with an even number of numerical points if we want to have no middle point but, rather, wish to classify, unequivocally, each and every student performance as either "acceptable" or "unacceptable." We have recommended in the next chapter "Performance Assessment" that for the purposes of local assessment in Illinois the even scale is the better choice.

Selecting Raters for Performance Ratings

Raters for performances must possess a number of special attributes:

1. *Objectivity* (ability to rule out feelings)
2. *Knowledge of the "subject matter" of the academic or extra-curricular field*
3. *Trainability* (ability to profit from special training programs, workshops, practica, seminars, etc.)
4. *Freedom from the common biases (or errors) of rating scale use* (such as "halo effect", "generosity error", favoritism, etc.)
5. *Practical experience in using rating methods.*

Selection should take, at least, the first four attributes into account, with the proviso that #5 could be achieved under conditions of formal training.

Training Raters for Performance Ratings

The supervisor of a performance rating scale system would never simply hand a rating scale sheet to a new rater and tell him to go out and start rating. Even when a

rating scale is well written and well edited, we must determine if all raters interpret the descriptions and instructions in a uniform manner. When they don't, we need to do two things: (1) revise the instrument and (2) train the raters.

Initially, you should explain to your raters what you mean by the systematic observation required on a rating scale and to describe the behavior they will be watching. Raters should be trained in the definitions you will be using to describe the behavior, and materials and events which should be taking place. Copies of the rating instrument should be distributed and you should explain how to use it. It is also a good idea to explain to the raters why you need the information and how it will be used.

We cannot emphasize too strongly how useful videotapes can be in the training of raters. In paper-and-pencil assessment of academic (or cognitive) achievement, we have a tangible product; not so with performance rating, UNLESS WE VIDEOTAPE! When we tape, we have the (tangible) tape to reproduce the images of the performance and can thereby repeat ratings by one rater at different points in time, or we can use the tape for training purposes.

The initial training should be sufficient to get raters started on the job. After they have functioned for a time, we will be able to distinguish whether several raters agree on their ratings of the same performances or whether some raters deviate a greal deal (some "tough", others "too easy"). If the deviations are too large among raters, then a second training session is called for. The purpose of the second training session will be to get at the nature of disagreements, to bring these to the attention of the raters, and help raters to reduce the variation in ratings.

Common Errors in Rating

Before getting into specific recommendations for training methods, let us look in more detail at some of the common so-called "errors" of raters.

The next few paragraphs are a quotation, verbatim, from two pages in one of the References of this Guide (Gronlund, 1985, pp.396-397).

> Certain types of errors occur so often in ratings that special efforts are needed to counteract them. These include errors due to (1) personal bias, (2) halo effect, and (3) logical error.
>
> *Personal bias* errors are indicated by a general tendency to rate all individuals at approximately the same position on the scale. Some raters tend to use the high end of the scale only, which is probably the most common type of bias and is referred to as the *generosity error*. Occurring much less frequently but persistently for some raters is the *severity error*, in which the lower end of the scale is favored. Still a third type of constant response is shown by the rather who avoids both extremes of the scale and tends to rate everyone as average. This is called the *central tendency error*. It also occurs much less often than the generosity error does, but it tends to be a fixed

response style for some raters.

The tendency of a rater to favor a certain position on the scale has two undesirable results. First, it puts in doubt a single rating of an individual. A high or low rating might reflect the personal outlook of the rater rather than the personal characteristics of the person rated. This is not quite so serious, if the ratings are used only in the local school setting. In a local school situation, we are apt to know the rating habits of individual teachers and are thus able to discount their tendencies to overrate or underrate. Second, favoring a certain position on the scale limits the range of any individual's ratings. Therefore, even if we make allowances for a teacher's general tendency to rate pupils high, the ratings for different pupils may be so close together that they fail to provide reliable discriminations.

The *halo effect* is an error that occurs when a rater's general impression of a person influences the rating of individual characteristics. If the rater has a favorable attitude toward the person being rated, there will be a tendency to give high ratings on all traits, but if the rater's attitude is unfavorable, the ratings will be low. This differs from the generosity and severity errors in which the rater tends to rate everyone high or everyone low.

Because the halo effect causes a pupil to receive similar ratings on all characteristics, it tends to obscure strengths and weaknesses on different traits. This obviously limits the value of the ratings.

A *logical error* results when two characteristics are rated as more alike or less alike than they actually are because of the rater's beliefs concerning their relationship. In rating intelligence, for example, teachers tend to overrate the intelligence of pupils with high achievement because they expect the two characteristics to go together. Similarly, teachers who hold the common but false belief that gifted pupils have poor social adjustment will tend to underrate them on social characteristics. The errors, here, do not result from biases toward certain pupils or certain positions on the rating scale, but rather from the rater's assumption of a more direct relationship among traits than actually exists.

The various types of errors that appear in ratings are rather disconcerting to the classroom teacher who must depend on rating scales for evaluating certain aspects of learning and development. Fortunately, however, the error can be markedly reduced by proper design and use.

A systematic plan for the "second session" for training raters has been given in one of the References included in the Annotated Bibliography (p.356, Berk, 1986). It consisted of a nine-step training program for performance rating:

1. *A videotape of a job being performed is shown.*
2. *Trainees rated the performance on rating scales provided.*
3. *Each trainee's ratings were posted on a flip chart.*
4. *Differences and reasons for the differences in ratings were argued in group discussion.*
5. *Trainer gave the "correct" ratings that illustrate avoidance of the rating error being studied (whether similar-to-me, halo, or contrast error).*

6. *Trainer explained the "correct" ratings in terms of the ratee behavior that defined them.*
7. *Trainer discussed the general nature of the error being studied.*
8. *Trainees provided examples (in group discussion) of that kind of error in their own experiences.*
9. *Group discussed ways to overcome error.*

Although only three kinds of errors were treated in the research on which the nine steps were based, it was pointed out that "a variety of different cognitive errors could be treated by a similar procedure."

It is to be expected that raters who do no rating over a period of weeks or a period of months will get "rusty." Re-training sessions, which might well be called "re-calibration" of raters will be needed. We recommend including in any re-calibration sessions relevant aspects from the initial and the second session in the foregoing, so that the recalibration session would tend to run for a shorter period of time. Further, re-calibration can be expedited if you have "archived" anchor taped performances or products from the past. Those can serve as "practice" before actual rating in a subsequent year.

Single-Rater Reliability

It is possible that only one rater is available. Furthermore, we may be interested in this one rater's consistency of ratings over time. Without another rater or some of last year's ratings available, this is probably the only way to assess this one rater's stability in making ratings. The rater would need to repeat his/her rating at least once (total of two ratings per rater). It is desirable to space the two ratings over a period of time (at least one day, but preferably two to seven days). In this way we gather information on whether the rater changes over time (undesirable) and how much the rater changes. The more the rater changes the less accurate (or reliable) the rater is.

In the worked-out numerical examples of various indexes in the Exhibits, this case of the single-rater making two ratings at two points in time is not included. The reason is that the method and arithmetic which should be used is exactly the same as that for two-raters each making a single rating at the same point in time. Therefore, in the two-rater examples, simply substitute a single-rater's pair of rating scores for each student for the two raters' pair of scores for each student and then proceed as shown in the Exhibits.

Two-Rater Reliability

In the condition of two raters, we have them both do their ratings on the same performances at the same time. We are interested in determining how closely their ratings agree, but there are two ways in which they could agree: (1) They could agree in their rankings of students or (2) They could agree in the absolute numerical size of their rating scores. Both of the deviations from agreement (the opposite of agreement) are important and both should be considered.

Two methods of establishing the consistency of two raters of the same performances for a number of students are shown in worked-out numerical examples in Exhibit C (Case 1 and Case 2). The 40 rating scores in the data boxes for both cases are identical.

Step 1--Exhibit C

In Case 1, having copied the 40 scores into the data box, we add down each of the columns labelled X and Y and obtain the column totals, 57 and 48. Next we subtract Y from X for each of the 20 students and record these differences in the (Diff) column. Square each difference and record the square in the $(Diff)^2$ column. Add the X and Y scores for each student and record the sums in the (X+Y) column. Square each of these sums and record the square in the $(X+Y)^2$ column. Of course, you should check your arithmetic at every opportunity. For example, check to make sure that the column totals for columns X and Y (57 & 48) add up to 105.

Case 1: EXHIBIT C-- Computing Diagram For Reliability of Two-Raters of Multiple Student

STEP 1

Student No.	RATERS X	RATERS Y	(X-Y) Diff	$(X-Y)^2$ $(Diff)^2$	(X+Y) Student Total	$(X+Y)^2$ $(Student\ Total)^2$
1	3	3	0	0	6	36
2	3	3	0	0	6	36
3	3	3	0	0	6	36
4	4	3	1	1	7	49
5	3	3	0	0	6	36
6	1	1	0	0	2	4
7	3	3	0	0	6	36
8	5	4	1	1	9	81
9	4	3	1	1	7	49
10	4	3	1	1	7	49
11	4	3	1	1	7	49
12	4	3	1	1	7	49
13	2	1	1	1	3	9
14	2	1	1	1	3	9
15	1	1	0	0	2	4
16	2	1	1	1	3	9
17	1	1	0	0	2	4
18	3	3	0	0	6	36
19	3	3	0	0	6	36
20	2	2	0	0	4	16
TOTALS	57	48	9	9	105	633

Step 2

Two terms are needed for the Reliability Equation: A and B. To calculate A, first divide the column total for column $(Diff)^2$ (9) by the number of students (20) and subtract from this the square of the quotient of the column total for the (Diff) column divided by the number of students. Thus 9 divided by 20 is .45 and if we subtract from this the square of .45 we obtain .25 for A. To obtain B, we divide the sum of the $(X+Y)^2$ column (633) by the number of students (20), yielding 31.65. From this we subtract the square of [105 (sum of column "X+Y") divided by 20 (number of students)], which is 27.56. The subtraction yields 4.09 for B. Reliability is simply the difference between 1 and A/B (.06), and the final figure for Reliability is equal to .94.

EXHIBIT C (cont)

STEP 2 Terms for Reliability Equation

$A = 9/20 - (9/20)^2 = .45 - .20 = .25$

$B = 633/20 - (105/20)^2 = 31.65 - 27.56 = 4.09$

Reliability = 1 - .25/4.09 = 1 - .06 = .94

In Exhibit C (Case 2) the data from Case 1 have been copied into a new data box. In Step 1 a column labeled "Difference" is added to the right of the Data Box. Entries in this column are the difference between the first and second raters (X-Y). In the data shown all the differences are either "0" or "1," which is mere coincidence; in many cases the differences would be 2 or more.

Case 2: EXHIBIT C (cont) Rater Agreement of Two Raters of Multiple Students

STEP 1

Student No.	RATERS X	RATERS Y	Difference (X-Y)
1	3	3	0
2	3	3	0
3	3	3	0
4	4	3	1
5	3	3	0
6	1	1	0
7	3	3	0
8	5	4	1
9	4	3	1
10	4	3	1
11	4	3	1
12	4	3	1
13	2	1	1
14	2	1	1
15	1	1	0
16	2	1	1
17	1	1	0
18	3	3	0
19	3	3	0
20	2	2	0

In Step 2 a simple index of agreement has been applied to our differences. Results of applying this index to various rating scale components (such as Integration, Focus, and Organization) are reported in tables labeled, "Percentage of Inter-rater Agreement Levels within One Point for Each Feature."

EXHIBIT C (cont)

STEP 2 Rater Agreement (%) =

100 [No. of Differences Less Than 2]/[Number of Students Rated]
100 [20/20] = 100 (1) = 100%

We have chosen to call the index "Rater Agreement (%)." The formula is the ratio of the "No. of Differences Less Than 2" divided by the Number of Students and the ratio multiplied by 100 yielded and index value of 100% for the Rater Agreement (%).

No guidelines are available for interpreting the Rater Agreement (%).

Multi-Rater Reliability--Exhibit A

In our indices of reliability of multiple raters, we are interested in establishing how much consensus there is among a set of rater's scores. (1) One component represents the differences (or fluctuations) among the raters. (2) Another component represents the chance variation or random error, which is beyond human control. The greater the chance or random effects, the less consensus there is among the raters; the less the chance or random effects, the greater consensus there is among the raters.

So much for theory. The arithmetic for computing a measure of consensus (or reliability) among raters is given in Exhibit A.

Begin Exhibit A, Step 1, by setting up a blank with Students down the rows and Raters across the columns. Enter each rating score in a "cell" of the table at the intersection of a particular student rated by a particular rater. For example, Student no. 9 was rated "5" by Rater C. Fill in all of the cells (5x20=100 in this case); we are calling this set of all the original scores a "Data Box." Now, compute the row totals (here called "Rater Total") and column total (here called the "Student Totals"). For example, adding across for the 10th student yields a row total of "20". Adding down the third column yields a column total of "69." The number "313" is the sum of the Rater Totals and it is also the sum of the Student Totals. This is called a "Check Sum", and you always must be able to get the same Check Sum by adding first one set of totals and then the other set of totals.

EXHIBIT A--Computing Diagram for Reliability of Multiple-Raters of Multiple-Students

STEP 1 Rating Scores and Totals for Original Data

Student No.	RATERS A	B	C	D	E	Student Totals
1	3	3	3	3	4	16
2	3	3	3	3	5	17
3	3	3	4	3	4	17
4	4	3	4	3	5	19
5	3	3	4	3	4	17
6	1	1	2	2	3	9
7	3	3	3	3	4	16
8	5	4	5	4	5	23
9	4	3	5	4	5	21
10	4	3	4	4	5	20
11	4	3	5	4	5	21
12	4	3	4	4	5	20
13	2	1	3	3	3	12
14	2	1	3	3	3	12
15	1	1	2	2	3	9
16	2	1	2	2	3	10
17	1	1	2	1	2	7
18	3	3	4	3	4	17
19	3	3	4	4	4	18
20	2	2	3	2	3	12
Rater Totals	57	48	69	60	79	313

Go to Step 2 and set up a new "box." The new box of 100 cells has, as entries, the square of each of the original scores in Step 1. Since all the scores (from 1 to 5) are small, and since everybody knows the squares of these numbers, you can simply glance at any score in the Data Box of Step 1, instantly square it in your head, and quickly write it down in the corresponding cell in the Data Box in Step 2. Now we are ready to "sum the squares." There are three such sums needed. The first is the Sum of Squares of Rater Totals. Pick up each of the 5 Rater Totals, in turn, square it and record the square in the corresponding position at the foot of the appropriate column. For example, the square off "69" is 4,76l and that number has been recorded at the foot of the third column in Step 2. When you have squared all the Rater Totals, add them up, yielding in our example the number 20,155. However, hold on for just a moment! We don't expect you to square a number like "69" in your head (or for that matter with pencil and paper). It's just too easy nowadays to do such squares on a hand calculator (even on the most inexpensive model now selling for $4.99 or less). All you need do is to enter the number to be squared (say, 69), press the "X" (times) key and the "=" key in that order and simply read out the square of 69 in the display as "4,76l." Any other square can be obtained in the same way.

Exhibit A (cont)

STEP 2 Squares of Rating Scores and Totals

Student No.	RATERS A	B	C	D	E	Squares of Student Totals
1	9	9	9	9	16	256
2	9	9	9	9	25	289
3	9	9	16	9	16	289
4	16	9	16	9	25	361
5	9	9	16	9	16	289
6	1	1	4	4	9	81
7	9	9	9	9	16	256
8	25	16	25	16	25	529
9	16	9	25	16	25	441
10	16	9	16	16	25	400
11	16	9	25	16	25	441
12	16	9	16	16	25	400
13	4	1	9	9	9	144
14	4	1	9	9	9	144
15	1	1	4	4	9	81
16	4	1	4	4	9	100
17	1	1	4	1	4	49
18	9	9	16	9	16	289
19	9	9	16	16	16	324
20	4	4	9	4	9	144
Squares of Raters	3249	2304	4761	3600	6241	5307

SUM OF SQUARES OF RATER TOTALS = 20155

The next sum of square to be calculated in the Sum of Squares of Student Totals. Pick up, in turn, each Student Total from Step 1, square it and record the square on the appropriate line in the "Square of Student Totals" column in Step 2. Add up the column and obtain the number 5,307.

Exhibit A (cont)

STEP 3 Sums of Squared Scores and Totals

Sum of 20 Student Totals	= 313
Sum of 100 Squares of Student Ratings	= 1,101
Sum of 5 Squares of Rater Totals	= 20,155
Sum of 20 Squares of Student Totals	= 5,307

The third and final Sum of Squares is the sum of the squares in all 100 of the cells of Step 2, which comes out to the 1,101. Step 3 summarizes these three Sums of Squares and makes them readily accessible for substitution in Step 4.

Step 4 (Terms of Reliability Equation) records or calculates all the three terms necessary for the final reliability equation. Record the constant k=5, the number of raters. Calculate term A as follows. Divide the Sum of Squares of Student Totals (5,307) by 20, the number of students, giving us the number 265.35. Subtract from this number of the sum of the 100 scores in Step 1 (313) divided by the square of the number of students (202=400), the number 244,9225 and obtain the value 20.4275 for Term A. Calculate Term B as follows. Divide the Sum of Squares of Student Ratings (1,101) by 20 (the number of students), obtaining 55.05. From 55.05 subtract 20,155 (Sum of Square of Rater Totals) divided by the square of the number of students (400) to obtain 50.3875. The difference obtained by subtraction gives us 4.6625, Term B.

Exhibit A (cont)

STEP 4 Terms of Reliability Equation

n = 20, Number of Students

K = 5, Number of Raters

$$A = \boxed{\frac{\text{Sum of Squares of Student Totals}}{n}} - \boxed{\frac{\text{Square Sum of Student Totals}}{n^2}} \quad \boxed{\frac{5,307}{20}} - \boxed{\frac{97,969.00}{400}} = 265.35 - 244.9225 =$$

A = 20.4275

$$B = \boxed{\frac{\text{Sum of Squares of Student Ratings}}{n}} - \boxed{\frac{\text{Sum of Squares of Rater Totals}}{n^2}} = \boxed{\frac{1,101}{20}} - \boxed{\frac{20,155}{400}} = 55.05 - 50.3875 =$$

B = 4.6625

In Step 5 the terms k/(k-1) in our case is 5/4=1.25. In the brackets, B/A is 4.6225/20.4275=.2282. Next, 1-.2282=.7718. Finally, 1.25 multiplied by .77.8=.96, the Reliability for our 100 ratings of 20 students by 5 raters. This kind of reliability is frequently referred to as "Inter-Rater Reliability."

Exhibit A (cont)

STEP 5 Reliability Calculations

$$\text{Reliability} = \frac{k}{k-1}\boxed{1 - \frac{B}{A}} = \frac{5}{4}\boxed{1 - \frac{4.6625}{20.4275}} = 1.25\,(1 - .2282) = 1.25\,(.7718)$$

Reliability = .96

Worksheet

HERE'S A PROBLEM!

Up to now you have been passively observing our logical, but long, exposition of the rationale and the steps in calculating reliability coefficients. This isn't enough to complete your training. Now you need "HANDS ON" experience in carrying out the process from raw data to finished reliability coefficient.

Below is Table H (for "hands-on") showing proficiency ratings on a scale from 1 to 6 of 20 students by 3 judges on the specified dance exercise used in our videotapes.

Follow the "cookbook" in Exhibit A and experience the satisfaction of computing your very own reliability coefficient. Then compare it with Exhibit G.

Table H Dance Ratings Data

Student	Judge A	Judge B	Judge C	Student Totals
1	6	5	5	
2	3	3	2	
3	6	6	6	
4	5	3	5	
5	3	1	3	
6	4	4	5	
7	5	5	5	
8	3	3	4	
9	4	4	4	
10	3	3	3	
11	4	3	4	
12	3	2	3	
13	3	3	4	
14	4	3	4	
15	4	4	3	
16	1	1	2	
17	4	4	4	
18	2	1	1	
19	1	1	1	
20	4	4	4	
Rater Totals				

CHAPTER BIBLIOGRAPHY

ALLISON, DONALD E., "The Effect of Homogeneous vs. Heterogeneous Matching - Format on Test Performance and Reliability." The Alberta Journal of Educational Research, Vol. XXX, No. 2, June 1984, Pages 91-94.

BARKER, LARRY, WATSON, K.W., KIBLER, R.J., "An Investigation of the Effect of Presentations By Effective and Ineffective Speakers on Listening Test Scores." The Southern Speech Communication Journal, Spring 1984, Pages 309-318.

BASCH, CHARLES E. & GOLD, ROBERT S., "The Validity of Reliability Assessments." Journal of School Health, May 1985, Vol. 55, No. 5, Pages 191-195.

BERK, RONALD A., "Setting Passing Scores on Competency Tests." NASSP Bulletin, February 1987, Pages 69-76.

BOWER, RUTH, "Matching Standardized Achievement Tests to Local Objectives." ERS Spectrum, Vol. I, No. 1, Spring 1983, Pages 40-43.

CITRON, CHRISTIANE H., "Courts Provide Insight on Content Validity Requirements." Educational Measurement: Issues and Practice, Winter 1983, Pages 6-7.

DURAN, ROBERT L., "Communicative Adaptability: A Measure of Social Communicative Competence." Communication Quarterly, Vol. 31, No. 4, Fall 1983, Pages 320-326.

ENGLISH, FENWICK W., "The Principal as Master Architect of Curricular Unity." NASSP Bulletin, April 1987, Pages 35-42.

FISHER, THOMAS H., "Implementing an Instructional Validity Study of the Florida High School Graduation Test." Educational Measurement: Issues and Practice, Winter 1983, Pages 8-9.

FREDERIKSEN, NORMAN, "Construct Validity and Construct Similarity: Methods for Use in Test Development and Test Validation." Multivariate Behavioral Research, January 1986, Pages 3-28.

FRISBIE, DAVID A., "Reliability of Scores from Teacher-Made Tests." (NCME Instructional Module Series) Educational Measurement: Issues and Practice, Spring 1988, Pages 25-35.

GLATTHORN, ALLAN A., "Teacher Autonomy vs. Curricular Anarchy." NASSP Bulletin, April 1987, Pages 77-84.

GREEN, DONALD ROSS, "A Guide for Interpreting Standardized Test Scores." NASSP Bulletin, February 1987, Pages 23-35.

HAERTEL, EDWARD, "Construct Validity and Criterion-Referenced Testing." Review of Educational Research, Spring 1985, Vol. 55, No. 1, Pages 23-46.

KUMAR, V.K., REHILL, K.B., TREADWELL, T.W., & LAMBERT, P., "The Effects of Disguising Scale Purpose on Reliability and Validity." Measurement and Evaluation in Counseling and Development, January 1986, Pages 163-167.

MEHRENS, WILLIAM A., & LEHMANN, IRVIN J., "Using Teacher-Made Measurement Devices." NASSP Bulletin, February 1987, Pages 36-44.

NACCARATO, RICHARD W., "A Generalizability Analysis of a Rhetoric Experiment." Presented at the annual college conference on composition and communication, Washington D.C., Spring 1980.

NEVO, BARUCH, "Face Validity Revisited." Journal of Educational Measurement, Vol. 22, No. 4, Winter 1985, Pages 287-293.

POPHAM, W. JAMES, "Can High-Stakes Tests Be Developed At The Local Level?" NASSP Bulletin, February 1987, Pages 77-84.

RUDMAN, HERBERT C., "Classroom Instruction and Tests: What Do We Really Know About the Link?" NASSP Bulletin, February 1987, Pages 3-18.

SHOHAMY, ELANA, "The Stability of Oral Proficiency Assessment on the Oral Interview Testing Procedures." Language Learning, Vol. 33, No. 4, Pages 527-540.

SHOHAMY, ELANA, & REVES, T., "Authentic Language Tests: Where From and Where To?" Language Testing, Vol. 2, No. 1, June 1985, Pages 48-59.

STEVENSON, DOUGLAS K., "Authenticity, Validity and a Tea Party." Language Testing, Vol. 2, No. 1, June 1985, Pages 41-47.

ZIMMERMAN, DONALD, W., "The Mathematical Definition of Test Validity." Educational and Psychological Measurement, 1983,43, Pages 791-796.

PERFORMANCE ASSESSMENT

CHAPTER THREE

Many times in a school setting, teachers must assess a student performance or a student-made product (artifact). The guidelines for such assessment appear on the surface to be much more judgmental and arbitrary than for a paper and pencil test. Further, performance assessment data while frequently used in grading students in language arts, fine arts, physical education, and industrial arts is seldom reported as part of student achievement profiles.

PERFORMANCE ASSESSMENT: OBJECTIVES AND ASSIGNMENTS

In the first chapter, "Making Sure," we have addressed the issue of establishing good objectives. Performance assessment, too, begins with clear objectives and proceeds through focused instruction to appropriate assessment.

But, we're getting ahead of ourselves. Let's assume for a moment that you have responsibility for fine arts assessment. Look over the objectives for fine arts; are they in the performance domain or are they in the cognitive domain? Chances are you are answering, "both" which is not atypical. You are, however, going to have to determine an appropriate balance between the cognitive and performance domains. For example, at Homewood-Flossmoor speech is a required course for all students. The cognitive domain accounts for approximately 40% of a final grade, the performance domain counts for 60%, and no student may pass the course without at least one performance. In another example, the 1980-81 Connecticut Assessment of Educational Progress of Art and Music, at grades 4, 8 and 11, ten cognitive questions and two performance items are administered in art. Twenty-six cognitive questions and three performance items are used in music.

Characteristics of Good Assignments

After determining an appropriate balance, you must structure a valid and reliable assignment which will evoke typical performance levels for the performance objectives. You must carefully craft these assignments so that they are:

clear --every word should communicate
precise--a student should know exactly what is expected; a parent should know exactly what her child was required to do
apt --the assignment should fit the objective
reasonable --the assignment should be possible for students given the instruction they received
manageable --practicable, possible given the resources available

Process for Designing Assignments

If an assignment is part of your assessment, others should respond to it. Do they find it clear, precise, apt, reasonable, and manageable? As we developed the objectives and assignments for the training tape, we were especially concerned that each assignment could be understood by any teacher in the state regardless of training or teaching responsibilities. Five curricular teams of three learning area specialists designed the assignments. Teams then exchanged the assignments, went off to another area and read the other group's assignments. Next, they returned and performed the task as they understood it from their preliminary reading. You can imagine the fun we had as the dance group demonstrated a lay-up and vice versa. The assignments were then revised by the authoring group to further clarify the exact tasks.

We highly recommend a similar procedure whether you are part of a large curricular team or the only teacher at your grade level in your school. Give someone else a copy of the assignment you plan to use, have that person read it and perform the required task for you. You will quickly learn if you've been clear.

Certainly, if an objective states, "throws a ball a distance of ten feet" you may feel the objective is sufficiently clear. However, others may query: what size ball? thrown overhand or underhand? thrown with or without a step? In a school where two or more are teaching to that objective, the instruction itself may vary on each of the dimensions if the assignment is incompletely defined.

Our training tape includes short student performances of five skills:

throwing a 16" softball

shooting a lay-up

dancing: one combination

acting: characterizing old age

speaking: delivering an original speech

Examples

The following pages are the assignments used in preparing the tape.

THROWING A 16" SOFTBALL

OBJECTIVE: Demonstrate correct overhand throwing mechanics to a specified target using a 16" softball.

ASSIGNMENT: The student will throw a 16" softball to a target, using correct throwing mechanics.

The following description refers to a right-handed throw. (All descriptions reverse for a left-handed throw)

A. Grip during execution of throw
1. The ball is in the right hand.
2. The ball is held with the finger pads, fingers are apart and flexed around the ball.
3. The ball does not rest on the palm of the hand.

B. Preparation phase
1. Student faces the target, both feet pointed at the target and shoulder width apart. The ball is held with both hands, in front of the body, waist high. The right hand is gripping the ball as explained in Section A.
2. The right foot is shifted backward to a 90 degree angle to the target. There is a heel to heel relationship and the heels are shoulder width apart. The weight is now totally on the right foot.

3. The left arm and shoulder should face the target, the right shoulder is drawn back. There is upper body rotation. The left arm is straight and extended forward and to the ground.
4. The ball should be brought up to head height, slightly behind the right ear. The right elbow is shoulder high pointing backward.
5. Eyes are looking at target. The head is up.

C. Action Phase
1. The weight shifts from the right foot to the left foot. The left foot is lifted and should move 3"-6" toward the target.
2. As this weight shift is occurring, the ball is given forward impetus by shoulder rotation to the left.
3. The elbow should lead the arm action.
4. The ball is brought forward and released when the hand is slightly in front of the head.
5. Shoulders continue the rotation to the left.

D. Follow-Through Stage
1. The weight is on the left foot. The right foot has pivoted to the right bringing the heel up and the toe rests on the ground
2. The hips and shoulders are facing the target.
3. The right arm is extended towards the target. The right wrist is also extended.
4. The left arm is extended to the back and downward.

SHOOTING A LAY-UP

OBJECTIVE: Demonstrate a right handed lay-up using the correct mechanics.

ASSIGNMENT: The student will be able to demonstrate a right handed lay-up using the correct mechanics.

A. The Approach
1. The student will approach the backboard from a 45 degree angle
2. The student will carry the ball at chest height.
3. The student will protect the ball with both hands by grasping the ball on each side and slightly behind center. The fingers are relaxed and spread with thumbs pointing toward each other. The elbows will be extended slightly outward.
4. The student will pick a spot on or within the small square of the backboard on which he will lay up the ball.

B. The Take Off
1. The student will plant and take off with the left foot from a designated point.
2. The student will drive the right knee upward.
3. The student will simultaneously be moving the ball upward with both hands while rotating the ball so that he can see his knuckles on his right hand.
4. The student will continue to extend the right hand upward to full extension toward his target as his left hand drops from the ball.

C. The Release
1. The student will release the ball gently, letting the ball roll off his finger tips toward his target.
2. The student will remain looking at his target.

D. The Landing
1. The student will continue his follow through on the same 45 degree angle as in his approach.
2. The student will land slightly beyond the take-off point with his ankles and knees flexed.
3. The student will keep his focus forward on the follow through.

DANCING: ONE COMBINATION

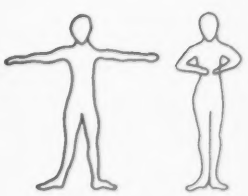
OBJECTIVE: Perform the following combination: *Balancé* right, *balancé* left, 3 step turn right, step left to open position.

ASSIGNMENT: The student will perform the following combination: *Balancé* right, *balancé* left, 3 step turn right, step left to open 4th position. This is to be done to 4 bars of 3/4 time music.

A. Start in 5th position
1. Right foot in front
2. Arms in low 5th position.

B. *Balancé* right: 3 steps (R,L,R) moving diagonally forward to the right.
1. lst step is diagonally forward on the right, in *demi-plié* with the emphasis on the lst step.
2. 2nd step, left foot closes to 5th position *relevé* behind the right foot.
3. 3rd step, right, is done in place, 5th position *relevé.*
4. Arms move to 3rd position, right arm above head on 1st beat.

C. *Balancé* left: 3 steps (L,R,L) moving diagonally back to the left.
1. lst step is diagonally back on the left in *demi-plié,* with the emphasis on the first step.
2. 2nd step, right foot closes to 5th position *relevé* in front.
3. 3rd step, left, is done in place, 5th position *relevé.*
4. Arms move to low 3rd position on the first beat.



D. 3 Step Turn to Right: 360 degree turn to right (R,L,R). End facing front. Entire turn is in *relevé.*
1. Arms: Right arm will move up through 1st position and open to 2nd position on 1st step.
2. On 2nd step, arms close to 1st position.
3. On 3rd step, arms open to 2nd position.

E. Step left to open 4th position.
1. On the first beat, step on left foot, *demi-plié,* crossing over right leg, diagonally, facing right front corner.
2. On 2nd beat, straighten left leg and point right toe back.
3. On beat 3 hold this position.
4. On 1st beat of this move, right arm goes through 2nd position to low 3rd position
5. On 2nd beat, right arm goes from low 3rd position to high 3rd position.
6. On 3rd beat, arms remain motionless.

ACTING: CHARACTERIZING OLD AGE

OBJECTIVE: Develop physical expressiveness which heightens student's ability to create old-age characterization.

ASSIGNMENT: Students will demonstrate old-age characterization with bodily movement and physical interpretation. The assignment requires control, consistency and expressiveness of body muscles while concentrating on a selected sensory skill.

Length of Assignment: 2 minutes. Two trials will take place.

Trial #1 - 20 seconds (used for baseline comparison)
Trial #2 - approx. 1:40 minutes (will be evaluated for skill demonstration)

I. Trial #1 (overview)

Seated in a chair, a student will hear a phone ring (imaginary stimulus) and will respond by rising and moving approximately 3 feet forward, picking up the receiver from telephone located on a table (imaginary stimulus) and say "Hello," identify self, and recognize the call as a wrong number. The student should replace the phone and return to the chair.

In the 1st trial, the following description will list the details of the acting task.

A. Student should recognize the sound of a telephone ring (by a facial response). The student should turn toward sound.

B. Student should visualize a spot where the telephone is.

C. Student must stand and then move 3 feet toward the phone (imaginary stimulus).

D. Student must pick-up receiver (with consistent emphasis on the size and weight of imagined object).

E. Student will respond by saying, "Hello," giving name, and stating that the call is a wrong number.

F. Receiver must be replaced to imagined level.

G. The student should return to chair in order to begin Trial #2.

(In Trial #2 the steps are re-created with the expanded instructional objectives of old age. (traits which are a "downward pull of gravity movements") Students will perform A-G listed below in a sustained physical manner.

II. Trial #2

A. The student begins in a seated position. All physical weight is pulled down and back as the student nearly slumps in the seat. Student head is down. Student recognizes the sound of a telephone ring. Head slowly raises in the direction of the sound. The student visualizes the spot where the phone is located.

B. The student stands slowly. Once standing, the weight of the body is placed backward so that the heels of the feet carry the weight.

C. The student walks toward the telephone. The heels must not leave the floor. The abdomen, stomach area, should remain slumped. The shoulders should also slump. The head should drop slightly with neck muscles relaxed. Knees should be slightly bent forward so that the weight appears to rest over the heel area.

D. The arm should extend up slowly toward the phone, drop slowly, and then pick up the receiver.

E. The student will sustain the old-age characteristics as the phone call (imaginary stimulus) is answered. The student will say, "Hello," give name, and state that the call is the wrong number. The student has the option to sound "older".

F. The receiver is replaced.

G. The student bows head to finish the task.

SPEAKING: DELIVERING AN ORIGINAL SPEECH

OBJECTIVE: Demonstrate flexibility in vocal and physical expression appropriate to the speaking situation.

ASSIGNMENT: Students will explain a personal incident that has affected them greatly. It could be an incident which caused them to experience much happiness, sadness, fear, uncertainty or any other emotion. This incident must be one that they feel they will never forget.

Assignment length: 2-3 minutes
Assignment must include the following:

A. Content
1. Introduction
   a. Captures audience's attention
   b. Establishes speaker-audience relationship
   c. Clearly indicates the topic.
2. Body
   a. Is organized chronologically
   b. Contains adequate details so the incident is understandable.
3. Conclusion
   Uses any appropriate type of conclusion (which has been previously taught).

B. Delivery
1. Facial expression must be consistent with the verbal message.
2. Ideas must be supported by appropriate gestures.
3. Speaker must be free from distracting mannerisms (i.e.swaying, pacing, head bobbing, etc.).
4. Voice must be varied (volume, rate, and inflection) to express ideas.
5. Eye contact with all areas of audience must be established during the speech.
6. Extemporaneous style of delivery must be used.
   a. A phrase outline is prepared in advance of speech, but exact wording must be left to the moment of delivery.
   b. One delivery card (index card) no larger than 4 inches by 6 inches and containing no more than 150 words written on one side only may be used.

RATING PERFORMANCE

We've already established that our project parameters are locally written and designed criterion-referenced tests. For our purposes here then, please be reminded that ultimately any scoring must indicate whether the student did or did not achieve the objective. Whether you plan to use holistic, analytic (trait), or combination scoring; or to assign ratings, ranks or use a checklist, you need to answer one question: *Has the student achieved the objective?*

Choosing the Right Criteria for Judging Performance

Raters must be able to define the standards they use to judge performance. These criteria should be:

--easily understood
--relevant
--compatible with one another
--precise
--representative of the vocabulary of the discipline
--objective
---unique, not overlapping

We feel the comments Patrick A. Byrnes of the Detroit Public Schools makes in his task description of the Creative Products Scales (p. 72) are particularly salient to this discussion. He explains the task force's goal:

> Accepting the fact that when working in the area of fine arts and creativity some form of human judgement must be brought to bear, can we, nevertheless, determine what kinds of information or criteria would most likely result in a more consistent, reliable and explainable evaluation of a student's creative product?
>
> The primary task is to develop a set of criteria which may be used to guide a panel of judges in determining the quality of a student's creative product.
>
> This set of criteria or rating scales will attempt to delineate the essential characteristics which, if observed in a creative product, would move a panel to judge its quality in a consistent manner, and would cause a second panel to arrive at approximately the same rating.
>
> If we tape-recorded one hundred three person panels as they discussed a creative product (in their respective area of the arts) what common remarks would emerge, what common characteristics would be mentioned time after time? These common characteristics, it would seem, are the criteria which the panel used for selecting one painting, poem, musical, dance or dramatic performance over another. Can we put these criteria into some kind of rating scale which could be used, anywhere, to determine whether a creative product shows that the student has acquired sufficient knowledge and skill to progress in a particular area of the fine arts?
>
> Would this set of criteria or rating scales enable a panel of judges to explain to a student whose creative product did not win a prize or was not selected for exhibition, why it did not win or was not selected?

The criteria which are developed should not forge new creative ground in the field. They should reflect qualities which are generally recognized to be present in superior performances of the skill.

After brainstorming a list of appropriate qualities within our subject groups, we exchanged lists and the groups met later to share their understanding.

Example Criteria

THROWING A 16" SOFTBALL: TRAITS

1. Grip
   Correct position of the ball in the hand.

2. Opposition
   Correct relative position of legs, body, and arms.

3. Transfer of weight
   Correct, balanced movement of the body: backward to forward shift.

4. Coordination
   Fluidity of movement.

5. Execution
   Correct movements in the prescribed sequence.

SHOOTING A LAY-UP: TRAITS

1. Mechanics
   The movement patterns the body goes through in performance of a skill
   a. Angle of Approach
      The student will approach the lay-up shot from a 45 degree angle to the backboard.
   b. Use of Arms and Hands
      1.) The student will move toward the basket carrying the ball chest high.
      2.) The student will simultaneously be moving the ball upward with both hands while rotating the ball so that he can see the knuckles on his right hand.
      3.) The student will protect the ball with both hands by grasping the ball on each side and slightly behind center. The fingers are relaxed and spread with thumbs pointing toward each other. The elbows will be extended slightly outward.
      4.) The student will continue to extend the right hand upward to full extension toward his target as his left hand drops from the ball.

2. Lift
   Leaving the floor with as much thrust as possible toward the target.
   a. The student will plant and take off on the left foot.
   b. The student will drive the right knee upward.

3. Accuracy

Ball goes through the net.

a. In Square

1.) The student will lay the ball up on a spot on or within the small square of the backboard.

2.) The student will remain looking at his target.

b. Force of Ball on Target

The student will release the ball gently, letting the ball roll off his finger tips toward his target.

DANCING: TRAITS

1. Execution

Correct perfor[illegible]nce of prescribed dance moves.

2. Placement

Posture and correct position of arms, head and feet.

3. Spatial Direction

Movement following designated floor plan.

4. Musicality

Match of motion to music using correct phrasing, tempo and rhythm.

5. Aesthetic Quality

Ease of movement and grace.

ACTING: TRAITS

1. Poise

Ease and self-assurance of manner in performing the activity

2. Body Movement

The coordination of movement reveals actions. In old age the movement is slow, weight is downward and back on heels of feet. Posture reveals shoulder weight; gestures lack energy.

3. Sustained Concentration

Often defined as focused attention on task, concentration must be maintained continuously. Actions are consistent with expectations; no break in character is evident.

4. Sensory Recognition

Imagined stimuli produce sensory responses in sight, sound and touch. Characterization tests belief in imagined task; the final result is physical expressiveness enriched by sensory detail imagined by student. (One observes in face and total body awareness the realization of the sensory imagery)

SPEAKING: TRAITS

1. The speech has an introduction, body, and conclusion
   a. The introduction
      1.) Contains an attention-getting device
      2.) Establishes a speaker-audience relationship
      3.) Orients the audience as to what will follow
   b. The body
      1.) Is logically organized
      2.) Contains effective, appropriate supporting materials
   c. The conclusion
      1.) Summarizes information or motivates the audience
      2.) Provides closure

2. Clarity of message
   a. Makes throughts and feelings understandable to audience
   b. Avoids obliqueness and ambiguity
   c. Is direct, specific, and honest
3. Quality of message (forethought)
   a. Is relevant to audience
   b. Is interesting to audience
   c. Is appropriate to audience
   d. Is sufficiently narrowed in scope
   e. Has a clearly defined thesis
   f. Is worded with personal and informal language
   g. Contains effective transitions between ideas

4. Effective use of body
   a. Appears natural (not phony) in terms of what audience knows about speaker
   b. Is appropriate in view of the audience, occasion, and subject matter
   c. Reinforces the message
   d. Does not call attention to itself

5. Expressive use of voice
   a. Aids intelligibility
   b. Sounds natural and comfortable (not phony and not strained)
   c. Emphasizes important ideas
   d. Is varied

RATING SCALES, CHECKLISTS, AND SCORING

Holistic vs. Analytic Scoring

A performance can be judged as a whole, according to one dimension, or according to a combination of dimensions. When the totality of the performance is the subject of global judgment, the scoring is termed holistic scoring. If the judgment is made on the basis of one main trait such as organization in composition, the scoring is termed primary trait scoring. Logically, secondary and tertiary trait scoring refer to the second and third characteristics which might be rated. The Rating Guide for Functional Writing used in Write on, Illinois!, Illinois State Board of Education, combines two approaches to evaluating student composition--holistic and analytic judgments.

You should only resort to trait scoring of a performance when you cannot get better information another way. If your objective is for your students to make a basket from the free throw line 75% of the time, you do not need trait scoring; you need to start counting hits and misses from among a number of shots. If the performance can be measured, measure it. Some examples of performances which can be measured are listed in the first chapter, "Making Sure." For your convenience we repeat the list here:

Keeping score at a game.
Using a stopwatch to time a sprint.
Using an audiometer to measure applause.
Using an emissions testing machine to check a tune-up.
Counting the bullet holes in each concentric ring of a bull's eye.
Using calipers to test the diameter of a rod.

Teresa Magnum, Director of the Writing Outreach Workshop at the University of Illinois at Urbana-Champaign shared The Freshman Rhetoric Manual authored by Dennis Baron which includes a clear set of composition elements which form the basis for writing evaluation. It is accompanied by a section titled "The Meaning of Grades". Both have been reproduced as a sample set of criteria and a prose form rating system.

The Six Elements of the Theme

Although there are many ways to analyze a piece of writing, we find it useful for instructional purposes to focus on six elements of the essay. These are listed on the back of each theme folder, and serve as a useful checklist for you when you revise and edit your essay before handing it in:

1. The writer's STANCE: Your thesis, or main point (is it easy for the reader to know what your essay is about? is your main point narrowed, specific, and focused? will your reader find your attitude acceptable or insulting?) Your role (your reader should know without having to guess just who you are and why you are writing) Your audience (have you thought about who you are writing to? what can you assume them to know already? and what new information do they need to know? is the tone and level of your writing

appropriate to the readers you address?)

2. CONTENT (is it accurate, specific, clear, appropriate?)
3. ORGANIZATION (do you proceed smoothly from start to finish? are the paragraphs and sections of your essay linked by appropriate transitions? are your purpose and conclusions clear to the reader? are your arguments supported by definite examples, facts or figures?)
4. SENTENCES (are they clear, logical, and varied? do they read well, or do they interfere with your reader's comprehension?)
5. DICTION (is your word choice specific, accurate, idiomatic? do you avoid clichés and bizarre or idiosyncratic expressions? are you pretentious or natural?)
6. MECHANICS (have you checked for standard punctuation, spelling, grammar, and format?)

These are only some of the elements that your instructor will develop with you during the course. They form the basis for our evaluation of your writing.

THE MEANING OF GRADES

Here are some guidelines that we use in classifying papers. Your instructor will explain in detail how he or she determines your grades.

A grade--To earn an A, a paper must be clearly superior, stating and developing its argument in a forceful, original manner, and showing that the writer is in complete control of his or her material. Such a paper exhibits no errors in mechanics, diction, or style.

B grade--This is a very good but not clearly outstanding paper. It lacks the level of originality and sophistication of an A paper. It is well organized, aimed at an apppropriate audience, and contains few if any errors in mechanics.

C grade--A satisfactory essay, neither distinguished nor severely flawed. It may contain some mechanical errors, but it need not. C is truly an average grade for a paper in Rhetoric courses.

D grade--A D represents poor work. It is a passing grade, though minimally so. It reflects a lack of mastery of two or more of the major elements of a theme, and generally shows a number of mechanical errors as well.

E grade--This is failing work. The essay in question is severely deficient in two or more of the elements of a theme. The mechanics may be glaringly inadequate. Alternatively, the essay may simply fail to fulfill the assignment you were given: it may be inadequate, incomplete or not on topic.

Although this system is used for grading rather than rating, notice how clearly the elements are defined and then joined with a quality judgment system, a grading scale.

Your rating scale should be:

functional --a teacher should be able to use it easily in the course of instructing.

understandable --a student, parent, or teacher should recognize precisely what the rating means

applicable to individual performance --a student's rating should be of his performance compared to the scale NOT to the performance of other students. In other words, these are ratings not rankings.

SOURCES FOR SAMPLE RATING FORMS

We queried a number of agencies and school districts for copies of rating scales already in use. It is impossible to share all of the materials we received. We feel these samples provide sufficient variety to prompt your thinking and your decision making.

Source: American Orff-Schulwerk Association
Sheran K. Fiedler, President
1406 Central Street
Evanston, Illinois 60201 (312) 475-4267

Resource:

The American Orff-Schulwerk Association provided a sample rating form used to evaluate children's performance groups auditioning for the 1987 National Orff - Schulwerk Association Conference at Hyatt Regency O'Hare, Nov. 11-15, 1987 (p. 96).

Source: Detroit Public Schools
Sharon Johnson, Evaluation and Testing
Detroit, Michigan (313) 494-2022

Resource:

The Detroit Public Schools developed a set of criteria which may be used to guide a panel of judges in determining the quality of a student's creative product. The scales were developed under the direction of Mr. Patrick Byrnes, Evaluation Project Director, Detroit Public Schools, and Dr. Beverly Ness Parke, Project Consultant, Wayne State University. The scales for judging art (p. 75), single dance (p. 76), and group dance (p. 78), drama (p.80), literature/drama (p. 83), music composition (p. 85), music performance (p.91), and speech (p.97) are included. The complete set is available through ERIC. Document: ED 218 903, EC 143 168 Author: Byrnes, Patrick A; and others Title: "Creative Products Scales--Detroit Public Schools" 36 p.; Paper presented at the Annual International Convention of the Council for Exceptional Children (60th, Houston, TX, April 11-16, 1982).

Source: Illinois High School Association
H. David Fry, Staff Administrator
2715 McGraw Drive, Box 2715
Bloomington, Illinois 61702-2715 (309) 663-6377

Resource:

The Illinois High School Association sponsors competition in a number of performance areas. Definitions and rules for each of the activities are described in separate rules books. One representative sample includes the section on Extempore Speaking from the 1987-88 IHSA Speech Rules Book (p. 99) and a copy of the IHSA Speech Critique for Individual Events, Dramatics, and Group Interpretation (p.100).

Source: Illinois Music Educators Association
6238-30 Columbus Street, Suite 203
Ottawa, Illinois 61350

Resource:

John Kinnison, Executive Director of the Illinois Music Educators Association, shared a number of nomination and audition procedures which are used in the selection of members of the IMEA performance groups. The Jazz Band Audition Procedure (p. 88) is a sample of weightings applied during the selection process. A number of forms provided from the Illinois Grade School Music Association, Northern Division (p.90) are also included.

Source: Illinois State Music Teachers Association
16706 Maple
South Holland, Illinois 60473

Resource:

The rating form for String Performance (p.87) was generously provided by Marguerite Rietvelt, Syllabus Coordinator of the Illinois State Music Teachers Association (ISMTA). ISMTA also has rating forms for rating both theory and piano performance and voice examination.

Source: Minnesota Department of Education
Capitol Square, 550 Cedar Street
Saint Paul, Minnesota 55101 (612) 296-4077

Resource:

Mark Youngstrom, Specialist in Communication Education, shared materials on the Minnesota Language Assessment: I984. One small set of forms relating specifically to narration are included as examples of the Minnesota materials (p.102).

Source: National Federation of State High School Association
11724 Plaza Circle, P.O. Box 20626
Kansas City, Missouri 64195

Resource:

The National Federation of State High School Association has forms available for adjudicating performances. The National Federation Music Adjudication Form--Solo (p. 94) is included as an example.

Source: Sourthern Illinois University
Department of Speech
Carbondale, Illinois 62901

Resource:

Dr. Marion Kleinau shared several rating forms for Oral Interpretation. She mentioned in her letter that she personally does not like to use such scales because she finds that it interferes with her ability to see the event as a whole and to pinpoint those particular strengths and weaknesses unique to that performance. Her vast experience and expertise in the field should not go underrated, however. One of the forms she sent is included (p. 101).

CREATIVE PRODUCTS SCALE--ART
DETROIT PUBLIC SCHOOLS

Work _____ Rater_____

Introduction: While it's relatively impossible to evaluate an art product with objectivity, the following group of elements and principles of art should transcend cultural, geographic, sex, and age barriers as grading tools. They are absolutes which may appear in total or in part of a painting, sculpture, or craft product regardless of the sophistication of the producer. Consequently, a Leonardo Da Vinci, a Papuan native, or a Cass Tech art student are all aspiring toward the same composite ideal of elements and principles while expressing their individual concepts.

The evaluator should be forewarned that the number 10, or unique category, is synonymous with masterpiece and should therefore be used with cautious reservation. A masterpiece can only be rendered by a master artist, the work has to be universally recognized, it must have a life of its own, be technically perfect, comprise past, present and future tendencies, and have staying power. Any additional comments you wish to make, please put on the back of this sheet.

PLEASE CIRCLE YOUR RATING

	DEVELOPING			ADEQUATE			COMPE-TENT		EXCEP-TIONAL	UNIQUE
1. UNITY	1	2	3	4	5	6	7	8	9	10
2. LINE	1	2	3	4	5	6	7	8	9	10
3. SHAPE	1	2	3	4	5	6	7	8	9	10
4. VALUE	1	2	3	4	5	6	7	8	9	10
5. VOLUME	1	2	3	4	5	6	7	8	9	10
6. FORM	1	2	3	4	5	6	7	8	9	10
7. TEXTURE	1	2	3	4	5	6	7	8	9	10
8. COLOR	1	2	3	4	5	6	7	8	9	10
9. CONTRAST	1	2	3	4	5	6	7	8	9	10
10. HARMONY	1	2	3	4	5	6	7	8	9	10

Average Rating =

GLOSSARY

1. UNITY -- Oneness, consistency or integration. Unity is reached when all the elements in a design work together harmoniously.
2. LINE -- An actual or implied mark path, mass or edge where length is dominant. The expressive quality of the line is determined by the sensitivity and skill of the artist.
3. SHAPE -- Shapes are areas enclosed by line, or differences of color, value or texture between a shape and the area around it.
4. VALUE -- Value is considered to be the lightness or darkness of a color or of a neutral gray. There are countless degrees of light and dark between the whitest white and blackest black.
5. VOLUME -- Denotes the space occupied by solid form or mass and also signifies defined regions of space, volume can refer to a solid like a pebble or to emptiness like a hole.
6. FORM -- The shape and structure of anything. A particular organization of shape capable of arousing the emotional and intellectual participation of the individual.
7. TEXTURE -- The actual or simulated representation of the physical characteristics of any particular surface.
8. COLOR -- The hue (actual color), chroma (intensity), and value (modification) as applied to the work of art.
9. CONTRAST -- Contrasting color as in complimentaries, contrast shapes as in circles and squares, contrasting texture as in rough and smooth, and contrasting form, such as organic versus geometric.
10. HARMONY -- The related scheme in a composition such as circles with circular lines or related color schemes such as analogous or monochromatic.

CREATIVE PRODUCTS SCALE--DANCE, SINGLE DANCER
DETROIT PUBLIC SCHOOLS

Performer_____ Rater_____

Judges: Use the ten point scale to rate each category. Please circle your rating.

9 - 10 = EXCELLENT
7 - 8 = GOOD
5 - 6 = ACCEPTABLE
3 - 4 = WEAK
1 - 2 = POOR

CATEGORIES	RATING									
Single Dancer Technique:	POOR			ACCEPTABLE				EXCELLENT		
1. Placement	1	2	3	4	5	6	7	8	9	10
2. Turnout	1	2	3	4	5	6	7	8	9	10
3. Elevation	1	2	3	4	5	6	7	8	9	10
4. Extension	1	2	3	4	5	6	7	8	9	10
5. Flexibility	1	2	3	4	5	6	7	8	9	10
6. Port de Bras	1	2	3	4	5	6	7	8	9	10
				Average Rating =						
1. Aesthetics and Musicality	1	2	3	4	5	6	7	8	9	10
				Average Rating =						
Physical Structure:	POOR			ACCEPTABLE				EXCELLENT		
1. Balance of Body	1	2	3	4	5	6	7	8	9	10
2. Feet and Legs	1	2	3	4	5	6	7	8	9	10
3. Back and Abdomen	1	2	3	4	5	6	7	8	9	10
4. Head and Neck	1	2	3	4	5	6	7	8	9	10
				Average Rating =						
Projection-Stage Presence	POOR			ACCEPTABLE				EXCELLENT		
1. Stage Presence	1	2	3	4	5	6	7	8	9	10
				Average Rating =						
				Average of Four Ratings =						

COMMENTS:

CREATIVE PRODUCTS SCALE--DANCE, SINGLE DANCER
DETROIT PUBLIC SCHOOLS

CRITERIA FOR RATING SINGLE DANCER

Single Dancer Technical Aspects

1. Placement
   - Weight forward over ball of foot
   - *Arm en bas (bras bas)* slightly ahead of shoulders
   - Spine straight (from hip line)
   - Shoulders relaxed-slighly forward (see arms)
   - Head erect, placed on top of neck (slightly stretched)
   - Eyes-eye level
   - Abdomen-lifted out of hips-presses against spine

2. Turnout
   - In hips--feet must follow degree of leg turnout in hip joint--pelvis must not be dropped or tilted.

3. Elevation
   - Body must remain quiet-in place. Arms do not assist. *Ballon* as important as height in air.

4. Extention
   - Hip of working leg pressed down-body lifted out of hip of base leg--shoulders and arm not involved.

5. Flexibility
   - Not to be confused with uncontrolled "loose jointedness".

6. *Port de Bras* and *Port de Corps*
   - Arms should be integral part of total body movement--if pelvis is dropped (tilted) and feet pronated, is it lazy muscles or a structural problem?

Aesthetics and Musicality

Suitability of music and wardrobe (if any) to dancer and Choreography.

Physical Structure

1. Length of legs and arms should balance body.
2. Muscle line should be long and easy.
3. Hips should not be wider than shoulders.
4. Neck, long with head well placed on top.
5. Waist long--hips should not be too close to ribcage.
6. Feet--toes approximately even, strong and lifted arch, high instep and ankle, no pronation, wide metatarsal.
7. Legs--well proportioned--if pronated or bowed, is it lazy muscles or structural?

Projection-Stage Presence

Is the dancer 'comfortable'--is there a natural projection?

CREATIVE PRODUCTS SCALE--DANCE, DANCE GROUP
DETROIT PUBLIC SCHOOLS

Choreographer_____ Title of Dance_____

Name of School_____ Director_____

Style_____ Length of Dance_____ _____ Number of Dancers_____

Theme of Dance_____ Rater_____

Judges: Use the ten point scale to rate each category. Please circle your rating.

9 - 10 = EXCELLENT

7 - 8 = GOOD

5 - 6 = ACCEPTABLE

3 - 4 = WEAK

1 - 2 = POOR

CATEGORIES	RATING									
Presentation:	POOR		ACCEPTABLE				EXCELLENT			
1. Technical Proficiency	1	2	3	4	5	6	7	8	9	10
2. Projection	1	2	3	4	5	6	7	8	9	10
3. Quality of Interpretation	1	2	3	4	5	6	7	8	9	10
4. Group Unity	1	2	3	4	5	6	7	8	9	10
5. Readiness to Perform	1	2	3	4	5	6	7	8	9	10

Average Rating =

Choreography:	POOR			ACCEPTABLE				EXCELLENT		
1. Inventiveness of idea	1	2	3	4	5	6	7	8	9	10
2. Choice of movement	1	2	3	4	5	6	7	8	9	10
3. Spacial design	1	2	3	4	5	6	7	8	9	10
4. Use of dynamics	1	2	3	4	5	6	7	8	9	10
5. Movement flow	1	2	3	4	5	6	7	8	9	10

Average Rating =

CREATIVE PRODUCTS SCALE--DANCE GROUP (cont)

Accompaniment:	POOR			ACCEPTABLE					EXCELLENT	
1. Suitability of choreography	1	2	3	4	5	6	7	8	9	10
2. Quality of use	1	2	3	4	5	6	7	8	9	10
3. Quality of recording	1	2	3	4	5	6	7	8	9	10
4. Quality of live performance	1	2	3	4	5	6	7	8	9	10

Average Rating =

Technical Aspects:	POOR			ACCEPTABLE					EXCELLENT	
1. Use of sets	1	2	3	4	5	6	7	8	9	10
2. Use of props	1	2	3	4	5	6	7	8	9	10
3. Quality of costuming	1	2	3	4	5	6	7	8	9	10

Average Rating =

Average of Four Rating =

COMMENTS:

CREATIVE PRODUCTS SCALE--DANCE, DANCE GROUP
DETROIT PUBLIC SCHOOLS

Guidelines for Rating Dance Groups

1. Presentation
   In considering public school groups, some consideration should be given to the length of time the group has been together.

2. Choreography
   Choreography chosen or designed by the group should not be weighed on the same scale as choreography chosen or designed by the director or teacher.

3. Accompaniment
   Instruments available under audition conditions can often alter the quality of recordings or live performance.

4. Technical Aspects
   If audition time limits do not permit full set-up examples of props, sets, wardrobe and description of light plot can serve, although judgment is difficult.

CREATIVE PRODUCTS SCALE--PERFORMING ARTS, DRAMA
DETROIT PUBLIC SCHOOLS

Performer_____ Rater_____

Audition cutting should be a 1-3 minute monologue from a published play, with an introduction not included within time limit.

Judges: Use the ten-point scale to rate each category. Determine final rating by adding the ratings of each category and dividing by the number of categories used. Terminology listed is to provide insight as to what is expected.

Two Creative Products Scales have been formulated. One for drama and one for speech. When evaluating selections in the area of oral interpretation, use the scale for drama. (Some items will not be applicable).

Each scale has four major categories to critique. They should be rated according to the following scale:

(Please Circle your rating)

9 - 10 = SUPERIOR

7 - 8 = EXCELLENT

5 - 6 = GOOD

3 - 4 = FAIR

1 - 2 = WEAK

CATEGORIES	RATING									
Voice:	POOR			ACCEPTABLE				EXCELLENT		
1. articulation	1	2	3	4	5	6	7	8	9	10
2. projection	1	2	3	4	5	6	7	8	9	10
3. phrasing	1	2	3	4	5	6	7	8	9	10
4. pronunciation	1	2	3	4	5	6	7	8	9	10
5. range	1	2	3	4	5	6	7	8	9	10
6. tone	1	2	3	4	5	6	7	8	9	10
7. inflection	1	2	3	4	5	6	7	8	9	10
8. volume	1	2	3	4	5	6	7	8	9	10
9. clarity	1	2	3	4	5	6	7	8	9	10

Average Rating =

PERFORMING ARTS, DRAMA (cont)

Movement:	POOR					ACCEPTABLE				EXCELLENT
1. blocking	1	2	3	4	5	6	7	8	9	10
2. gestures	1	2	3	4	5	6	7	8	9	10
3. facial expression	1	2	3	4	5	6	7	8	9	10
4. coordination	1	2	3	4	5	6	7	8	9	10
5. posture	1	2	3	4	5	6	7	8	9	10
6. poise	1	2	3	4	5	6	7	8	9	10
7. muscle tone/effective	1	2	3	4	5	6	7	8	9	10
8. use of body	1	2	3	4	5	6	7	8	9	10
9. spotting (Oral Interp Only)	1	2	3	4	5	6	7	8	9	10

Average Rating =

Characterization:	POOR					ACCEPTABLE				EXCELLENT
I. believability	1	2	3	4	5	6	7	8	9	10
2. concentration/ consistency	1	2	3	4	5	6	7	8	9	10
3. stage presence	1	2	3	4	5	6	7	8	9	10
4. sincerity	1	2	3	4	5	6	7	8	9	10
5. fresh	1	2	3	4	5	6	7	8	9	10
6. unique	1	2	3	4	5	6	7	8	9	10
7. creative	1	2	3	4	5	6	7	8	9	10
8. exciting	1	2	3	4	5	6	7	8	9	10
9. demonstration of emotion	1	2	3	4	5	6	7	8	9	10
10. spontaneity	1	2	3	4	5	6	7	8	9	10
11. confidence	1	2	3	4	5	6	7	8	9	10

Average Rating=

PERFORMING ARTS, DRAMA (cont)

Exposition:	POOR					ACCEPTABLE				EXCELLENT
1. understanding	1	2	3	4	5	6	7	8	9	10
2. material	1	2	3	4	5	6	7	8	9	10
3. accuracy of interpretation	1	2	3	4	5	6	7	8	9	10

Average Rating =

Average of Four Ratings =

COMMENTS:

CREATIVE PRODUCTS SCALE--LITERATURE, DRAMA
DETROIT PUBLIC SCHOOLS

CRITERIA FOR JUDGING DRAMA

Work _____ Rater _____

Rate each play according to these 10 criteria. Indicate to what extent the work meets the criteria according to this scale

1	2	3	4	5	6	7	8	9	10
To a very limited extent				Somewhat				To a great extent	

1. The play is well suited for production on a stage. _____
2. The play has a clearly felt "dramatic question." _____
3. The playwright shows skill in the use of internal and external conflict. _____
4. Action on the stage is employed effectively. _____
5. The play makes use of set and costuming. _____
6. The play is not limited to the private world of the playwright, but rather touches on some universal concern. _____
7. The dialogue reflects the character of the speaker.
8. The dialogue is blended with action and advances the plot. _____
9. The play makes skillful use of the form and structure of drama. _____
10. The characters are lifelike and interesting so that the audience is emotionally involved with at least one of them. _____

Total _____

COMMENTS:

EXPLANATION OF DRAMATIC CRITERIA

1. Suited for the Stage
   Plays set on life rafts or in forests work out better as films. This is also true of plays that depend on much action without dialogue.

2. Dramatic Question
   There should be a strong emotional element which holds the attention of the audience. Will he avenge his father? Will she get away with it? The more subtlety used in answering the question, the better. Plays with a weak dramatic question are likely to be dull.

3. Conflict
   Conflict is basic to drama. It draws us into the action. It should be latent in the characters and the plot should flow from it. Its skillful handling is the mark of a good playwright.

4. Action
   The characters should move about on the stage. Exits and entrances can establish secondary scenes. There should be stage business to keep the characters visually alive. Physical contact should be well motivated.

5. Set and Costume
   Imaginative use of details and a visual focal point are useful in a set. If a bare stage is used, there should be some reason for it.

6. Universal Concern
   The play should draw us into an as-if-real experience. We should be drawn into it emotionally and not have to ask, "What is this play about?" Somehow it should illuminate life.

7. Dialogue and Character
   Lines should reflect character so that we should be able to tell who is speaking in most cases. Dialogue should also expand our knowledge of the character so that we know him better as the play develops.

8. Dialogue and Plot
   The plot should be moved forward by the dialogue. The use of a narrator should be minimal. Preaching should be limited or non-existent.

9. Structure of Drama
   Such structural elements as complication, climax and denouement, building tension within a scene and the contrast of one scene with another all add to the excellence of a play. There should be an impression of unity and coherence created by the work as a whole.

10. Lifelike Character
    The audience needs to care about one or more of the characters. If he is a hero we should root for him, if he is a villain, we should desire his undoing.

CREATIVE PRODUCTS SCALE--MUSIC COMPOSITION
DETROIT PUBLIC SCHOOLS

Composer _____ Rater _____

Use the ten-point scale to rate each major category. Determine final rating by adding the ratings of each category and dividing by the number of categories used.

9 - l0--Superior; the highest level of creativity and technique. Original, innovative, consistent and mature in melodic, motive, harmonic, rhythmic, and formal expression and usage. Construction seems right and necessary in the sense that it cannot be improved upon, mechanical aspects correct notation appropriate.

7 - 8--Excellent; a very good composition. Some errors or weaknesses. Original and mature in most musical aspects. Accurate in most technical aspects.

5 - 6--Competent; a good composition. Several errors and weaknesses; lacking in one or more of the major categories.

3 - 4--Adequate; a fair composition. Many obvious errors and basic weaknesses. Little innovation or originality. Numerous technical errors of notation, transposition, instrumentation.

1 - 2--Poor;, a developing level of musical composition, indicating much needed improvement. Inconsistent and immature in melodic, motivic, harmonic, rhythmic and formal expression and usage. Musical ideas trite or contrived. Mechanical aspects incorrect; notation inappropriate.

Read the explanatory material before attempting to rate. (Please circle your rating.)

Functional Processes	RATING									
1. LINE	1	2	3	4	5	6	7	8	9	10
2. HARMONY	1	2	3	4	5	6	7	8	9	10
3. RHYTHM	1	2	3	4	5	6	7	8	9	10
4. FORM	1	2	3	4	5	6	7	8	9	10
5. TEXTURE	1	2	3	4	5	6	7	8	9	10
6. COLOR	1	2	3	4	5	6	7	8	9	10
Technical Processes										
7. NOTATION	1	2	3	4	5	6	7	8	9	10
8. INSTRUMENT USAGE	1	2	3	4	5	6	7	8	9	10

Average Rating =

COMMENTS:

How to Use the CREATIVE PRODUCTS RATING
SCALE for MUSIC COMPOSITION

This Creative Product Rating Scale was developed to evaluate musical compositions by gifted high school music students.

The rating scale consists of eight rather broad categories that relate to the musical and technical aspects of musical composition. Carefully read the explanation for each category before beginning to rate. Try to keep as many specific items in mind as possible. Remember that the task is to rate these elements simultaneously in terms of the whole composition using the complete ten-point rating scale described on the rating sheet.

Note that not all of the items are appropriate to all compositions. They are meant to be exemplary and used as appropriate.

Use the space at the bottom of the rating sheet to comment on specific aspects of the compositions and to provide constructive criticism for the student.

Each student should submit his/her composition with a complete score and a taped performance.

FUNCTIONAL PROCESSES

1. Line
   Motivic/method content and development, theme transformation, phrasing thematic relationships.

2. Harmony
   Tonal scheme, harmonic rhythm, progression, dissonance and coloristic usage, cadential effects and usage.

3. Rhythm
   Patterns (regular, irregular), combination of figures, meter, tempo.

4. Form
   Unity and contrast, material developed appropriately, form fits material, proper pacing of formal sections, expressive integration and development of the elements.

5. Texture
   Interaction of voices (polyphony, homophony), density, sonority.

6. Color
   Instrumentation (complementary, contrasting), dynamics (sforzandi, climax, dynamic curve).

TECHNICAL PROCESSES

7. Notation
   Clear and readable, appropriate for composition (traditional vs. contrived), markings proper and complete (articulation, bowings, tempo, phrasing, dynamics), transpositions correct.

8. Instrument Usage
   Parts playable, ideomatic, appropriate, well written.

LEVEL **ISMTA STRING PERFORMANCE EXAMINATION RATING SHEET**

Teacher's Code _____

STUDENT'S NAME _____

PLACE _____ Room_____

DATE _____ TIME _____

Student's Program

	Composer	Title of Piece
1		
2		
3		
4		
5		

A. ACCURACY (25 Points)---------- Points

	1	2	3	4	5
Notes					
Rests					
Note Values					
Bowing/Pizz.					
Intonation					
Fingering					

B. MUSICIANSHIP (32 Points)---------- Points

	1	2	3	4	5
Phrasing					
Rhythm					
Steady Tempo					
Technique					
L.H. Position					
R.H. Position					
Dynamics					
Tone Quality					
Oral Questions*					

C. MEMORY (8 Points)---------- Points

Remarks:

Essay or
5 Questions
(Levels V-X
Total---------- Points

Performance
Total---------- Points

Technique (played
and written)
Total---------- Points

Sight Reading/
Rhythm/Vocabulary
Total---------- Points

_____
Judge's Signature

GRAND TOTAL---------- Points

*Not applicable to Level I or to Level X students who are scholarship candidates.

IMEA Form JAZZ BAND AUDITION PROCEDURES

Senior Division

These audition procedures are provided to help ensure that the methods of evaluating jazz band selectees are effective and consistent throughout the state. These guidelines should be followed for both District and All-State jazz auditions. It is important that the District Presidents, Jazz Chairman, Audition Chairmen and all auditioners are familiar with these procedures.

The audition will consist of three parts: performance of required numbers, sightreading and improvisation. There is a possible total of 80 points.

The purpose of the audition is to determine proper seating order in the sections. If there is more than one jazz band to be selected, put the best players in the top band; don't attempt to make the bands equal.

The suggested point spread for each category is: Excellent 8-10; Good 6-7; Fair 4-5; and Poor 0-3. The final score should be totaled, and then the auditioner must enter the student's rank within his section of like instruments. The auditioner must sign and date the audition form.

JAZZ WINDS -- Procedures

A. Required Number - 40 possible points

1. Pick three excerpts to have the student play, preferably from the 1st part. If the student wants to audition on a lower part, he may.
2. The adjudicator will evaluate: part accuracy, technical facility, rhythmic accuracy and interpretation.
3. Possible ten points for each category -- 40 total.

B. Sightreading - 10 possible points

1. A special sightreading sheet will be given each adjudicator at the time of the audition. This sheet will be previously prepared by the Jazz Division Vice-President or his designee. In order not to compromise the sightreading test, this material will be a controlled item and will not be distributed to the adjudicators until the time of the auditions.
2. Correct jazz style, as well as rhythmic accuracy, should be judged.

C. Improvisation - 10 possible points

1. Use the James Aebersold record, Vol. I, Fifth Edition, Side 2, Band 2, F Blues.
2. A good quality record player or tape player is necessary!
3. Let each student play two or three choruses.
4. Do not use student rhythm section for improvisation audition.

D. Tone Quality and Intonation - 20 possible points

1. Enter scores for these categories last, after hearing all other parts of the audition.
2. Possible ten points for each category -- 20 total.

JAZZ RHYTHM -- Procedures

All rhythm instruments should audition separately and not as a part of a small combo of rhythm instruments. It would be best to have a qualified judge for each of the different rhythm instruments, rather than one person to judge all of them. Most of the audition procedures listed for jazz winds will also apply to jazz rhythm instruments, with these exceptions:

IMEA Form (cont)

PIANO AND BASS:

1. Required number - 40 possible points
2. Sightreading - 10 possible points
   a. Sightreading test should include both written out parts and chord symbol parts.
3. Improvisation - 10 possible points
   a. Pianists should play self-accompanied solos, blues progression in the keys of Bb and F. Bass should also play alone.
   b. Two choruses in each key.
4. Rhythmic Stability - 20 possible points
   a. To be judged on the total overall audition.

GUITAR AND VIBES:

1. Required Number - 40 possible points
2. Sightreading - 10 possible points
   a. Include both written out parts and chord symbol parts.
3. Improvisation - 10 possible points
   a. Guitar and vibes should follow same procedure as jazz winds.
4. Rhythmic Stability - 20 possible points
   a. Demonstrate ability to comp in both the swing and rock styles.

DRUMS:

1. Required Number - 40 possible points
2. Sightreading - 10 possible points
   a. Test should include at least two styles: Swing and Rock or Latin.
3. Improvisation - 10 possible points
   a. Play an original jazz drum solo, 16 measures, medium fast.
4. Rhythmic Stability - 20 possible points
   a. Demonstrate: jazz beat - medium and fast
      rock beat - medium
      brushes - medium swing

Revised May, 1985

IMEA Form

ILLINOIS GRADE SCHOOL MUSIC ASSOCIATION

NORTHERN DIVISION

Timpani Solo Contest

Judge's Comment Sheet

Name of Contestant _____ _____ Code Number_____

Instrument _____ Grade in School_____

Number of years played_____

Name of Selection

Order of Appearance _____Performance Time _____ Room\Number _____

(Judges Check Evaluation. Number 1 is Highest)

	1	2	3	4	GENERAL REMARKS
Intonation-Tuning of Timpani					
Tone-Correct playing area and stick appropriateness					
Technique-Rolls, sticking, and muffling					
Interpretation and Style					
Dynamics					
Rhythmic Accuracy					
Stage Presentation and Appearance					

TOTAL POINTS _____

Rating Code: 7-10 points, Div. I; II-I5 Points, Div., II; 16-28 Points, Div. III

COMMENTS ON PERFORMANCE
(Judge will please include suggestions for improvement)

OFFICIAL RATING:
DIV. I _____
DIV. II _____
DIV. III_____

Judge's Signature_____

CREATIVE PRODUCTS SCALE--MUSIC PERFORMANCE
DETROIT PUBLIC SCHOOLS

Performer _____ Rater _____

Use the ten-point scale to rate each major category. Determine final rating by adding the ratings of each category and dividing by the number of categories used.

9 - 10--Superior; the highest level of performance. Few errors or weaknesses.
7 - 8--Excellent; a very good performance. Some minor errors or weaknesses.
5 - 6--Competent; a good level of performance. Several errors or weaknesses; lacking in one or more of the major categories.
3 - 4--Adequate; a fair level of performance. Many obvious errors and basic weaknesses in several performance categories.
1 - 2--Poor; a developing level of performance. Weak and uncertain indicating much needed improvement. Numerous errors and weaknesses in all performance categories.

Read the explanatory material before attempting to rate. (Please circle before rating.)

	RATING									
1. TONE	1	2	3	4	5	6	7	8	9	10
2. INTONATION	1	2	3	4	5	6	7	8	9	10
3. TECHNIQUE	1	2	3	4	5	6	7	8	9	10
4. INTERPRETATION	1	2	3	4	5	6	7	8	9	10
5. IMPROVISATION (If Applicable)	1	2	3	4	5	6	7	8	9	10

Average Rating =

COMMENTS:

How to Use the CREATIVE PRODUCTS RATING SCALE for MUSIC PERFORMANCE

This Creative Product Rating Scale was developed to evaluate the overall performance attainments of gifted high school music students.

The rating scale consists of five rather broad categories. Carefully read the explanation for each category before beginning to rate. Try to keep as many specific and appropriate items in mind as possible. Remember that the task is to rate these elements simultaneously in terms of the whole performance (not segments of it), using the complete ten-point rating scale described on the rating sheet.

Note that not all items are appropriate to all types of performance. They are meant to be as broadly descriptive as possible to facilitate the use of this scale with instrumental/vocal, string/wind, and solo/ensemble performances.

Use the space at the bottom of the rating sheet to comment on specific aspects of the performance and provide constructive criticism for the student.

1. Tone
   Characteristic quality, control, balance, blend, vowels, breath support, ensemble.

2. Intonation
   Melodic line, chords, individual, with accompaniment, within sections, among sections.

3. Technique
   Fluency, fingering, (dexterity, even runs), accuracy (key, notes, rhythm), precision (attacks, releases), articulation (bowing, tonguing, consonents), pedal use.

4. Interpretation
   Phrasing, line, dynamics (shading, climaxes, subito), tempo (appropriate, steady), general style (staccato, legato, marcato), historical style (renaissance, baroque, classical, romantic), rhythm (accents, meter, patterns).

5. Improvisation
   Originality, variety, appropriateness to style, line compatible to accompaniment (chord progressions, scale).

CREATIVE PRODUCT SCALE--MUSICAL PERFORMANCE
ADDITIONAL FACTORS
DETROIT PUBLIC SCHOOLS

Performer _____ Rater_____

Use the ten-point scale to rate each major category. Determine the final rating by adding the ratings of each category and dividing by the number of categories used.

9 - 10--Superior

7 - 8--Excellent

5 - 6--Good

3 - 4--Fair

1 - 2--Poor

Read the explanatory material before rating. (Please circle your rating.)

RATING

	POOR			ADEQUATE					SUPERIOR	
1. MUSICAL SELECTION	1	2	3	4	5	6	7	8	9	10
2. GENERAL SPIRIT	1	2	3	4	5	6	7	8	9	10
3. STAGE DEPORTMENT	1	2	3	4	5	6	7	8	9	10

Average Rating =

COMMENTS:

How to Use the CREATIVE PRODUCT RATING SCALE
for MUSICAL PERFORMANCE--ADDITIONAL FACTORS

The purpose of this supplementary scale is to evaluate the music used and the extra-musical characteristics of the performance. (This evaluation is optional.)

1. Musical Selection
   Suitability of music (appropriate level of difficulty), musical value.

2. General Spirit
   Energy and drive, artistry, confidence, taste, enthusiasm, satisfaction, communication, genuineness, overall effect.

3. Stage Deportment
   Appearance, discipline, posture, attentiveness, effort, stage presence, relation to audience.

NFSHSA Form

NATIONAL FEDERATION MUSIC ADJUDICATION FORM
SOLO

_____ (Order of Appearance)

_____ (Program/Event No.)

Event _____ (tenor solo, trumpet solo, etc.) Class _____

School _____

Name of Soloist _____ Location-Contest/Festival _____

Selections	Composer	Publisher
1. _____	_____	_____
2. _____	_____	_____

Place one of these numbers in each circle below, then total carefully.

5 - A superior performance - outstanding in nearly every detail.
4 - An excellent performance - minor defects.
3 - A good performance - lacking finesse and/or interpretation.
2 - A fair performance - basic weaknesses.
1 - A poor performance - unsatisfactory.

AREAS OF CONCERN	COMMENTS
Tone Quality Consider: resonance, control, clarity, focus, consistency, warmth	
Intonation Consider: accuracy to printed pitches	
Rhythm Consider: accuracy of note and rest values, duration, pulse, steadiness, correctness of meters	
Technique (facility/accuracy) Consider: artistry, attacks, releases, control of ranges, musical and/or mechanical skill	
Interpretation, Musicianship Consider: style, phrasing, tempo, dynamics, emotional involvement	

NFSHSA Form (cont)

Diction - Vocal
Bowing - Strings
Articulation - Winds

Other Performance Factors
Consider: choice of literature, appropriate appearance, poise, posture, general conduct, mannerisms, facial expression (vocal), memory (if required)

Sight-Reading, Scales

TOTAL POINTS RATING COMPUTATION TABLE _____
(Signature of adjudicator)

Without Sight-Reading		With Sight-Reading
35-32 points	= Division I (Superior)	= 40-36 points
31-25 points	= Division II (Excellent)	= 35-28 points
24-18 points	= Division IIII (Good)	= 27-20 points
17-11 points	= Division IV (Fair)	= 19-12 points
10- 7 points	= Division V (Poor)	= 11- 8 points

Divisional Rating _____

NATIONAL FEDERATION
of State High School Associations
11724 Plaza Circle, P.O. Box 20626
Kansas City, MO 64195

AMERICAN ORFF-SCHULWERK ASSOCIATION

STUDENT PERFORMING GROUPS--EVALUATION FORM

NAME OF GROUP _____

GRADE LEVEL _____

DIRECTOR _____

KEY: 5 = excellent
4 = slightly above average
3 = average to below
2 = below average
1 = unacceptable
0 = not applicable

1. Singing (vocal quality, enunciation, vowel placement, etc.) ..... _____
2. Orff Instruments/Percussion ..... _____
3. Recorder Playing ..... _____
4. Movement ..... _____
5. Speech ..... _____
6. True Elemental Style ..... _____

COMMENTS:

Invite: YES_____ NO_____ NOT SURE_____

CREATIVE PRODUCTS SCALE--PERFORMING ARTS, SPEECH
DETROIT PUBLIC SCHOOLS

Performer_____ Rater_____

Speech should be 1-3 minutes in length with the style and topic a matter of your choice.

Judges: Use the ten-point scale to rate each category. Determine final rating by adding the ratings of each category and dividing by the number of categories used. Terminology listed is to provide insight as to what is expected.

Each scale has four major categories to critique. They should be rated according to the following scale:

(Please circle your rating)

9 - 10 = SUPERIOR

7 - 8 = EXCELLENT

5 - 6 = GOOD

3 - 4 = FAIR

1 - 2 = WEAK

CATEGORIES	RATING									
Voice:	POOR			ACCEPTABLE				EXCELLENT		
1. variety	1	2	3	4	5	6	7	8	9	10
2. expression	1	2	3	4	5	6	7	8	9	10
3. tone	1	2	3	4	5	6	7	8	9	10
4. rate	1	2	3	4	5	6	7	8	9	10
5. pitch	1	2	3	4	5	6	7	8	9	10
6. inflection	1	2	3	4	5	6	7	8	9	10
7. clarity	1	2	3	4	5	6	7	8	9	10
8. volume	1	2	3	4	5	6	7	8	9	10

Average Rating =

Diction:	POOR			ACCEPTABLE				EXCELLENT		
1. articulation	1	2	3	4	5	6	7	8	9	10
2. pronunciation	1	2	3	4	5	6	7	8	9	10
3. style	1	2	3	4	5	6	7	8	9	10

CREATIVE PRODUCTS SCALE--PERFORMING ARTS, SPEECH (cont)

4. dialect	1	2	3	4	5	6	7	8	9	10
5. acceptable language	1	2	3	4	5	6	7	8	9	10

Average Rating =

Delivery:	POOR			ACCEPTABLE					EXCELLENT	
1. eye contact	1	2	3	4	5	6	7	8	9	10
2. gestures	1	2	3	4	5	6	7	8	9	10
3. use of pause	1	2	3	4	5	6	7	8	9	10
4. emphasis	1	2	3	4	5	6	7	8	9	10
5. facial expression	1	2	3	4	5	6	7	8	9	10
6. sincerity	1	2	3	4	5	6	7	8	9	10
7. poise	1	2	3	4	5	6	7	8	9	10
8. posture	1	2	3	4	5	6	7	8	9	10
9. emotional appeal	1	2	3	4	5	6	7	8	9	10

Average Rating =

Content:	POOR			ACCEPTABLE					EXCELLENT	
1. timely	1	2	3	4	5	6	7	8	9	10
2. specific purpose	1	2	3	4	5	6	7	8	9	10
3. cohesive	1	2	3	4	5	6	7	8	9	10
4. use of evidence	1	2	3	4	5	6	7	8	9	10
5. organization	1	2	3	4	5	6	7	8	9	10
6. reasoning	1	2	3	4	5	6	7	8	9	10
7. local appeal	1	2	3	4	5	6	7	8	9	10
8. unique approach	1	2	3	4	5	6	7	8	9	10

Average Rating =

Average of Four Ratings =

COMMENTS:

IHSA 1987-88 SPEECH RULES BOOK
EXTEMPORE SPEAKING

1. Definition: Extempore Speaking is an event in which a student is given forty-five minutes to prepare an original speech indicating his knowledge of current events concerning an assigned topic.

2. Purpose: The purpose of Extempore Speaking is to encourage a student to gain a broad knowledge of current events and to develop the ability to analyze the significance of such events and prepare, in a short period of time, a meaningful speech which can be delivered in a skillful manner.

3. Presentation:
   a. Topics: Topics shall concern events which have been of state, national or international importance at any time between the beginning of the current school year and the date of the contest. They shall be worded in the form of questions which do not elicit blanket yes or no responses.
   b. Drawing of Topics: Contestants shall draw topics according to their order of speaking at intervals to provide each contestant a maximum of 45 minutes preparation. Contest managers shall publish and/or post specific preparation and performance times for each contestant in each round.

      Each contestant will be provided a packet of topics to be used throughout the contest. The contestant shall draw three topics and, within one minute, choose the one on which to speak. The choice shall be recorded by the monitor of the drawing. The two topics drawn but not selected will be returned to the packet. The topic chosen and spoken on will be discarded. Contestants will then select topics from the same packet in all rounds, but will speak on a different topic in each round.

      After all contestants have selected topics, a list of topics shall be prepared and given to the judge or judges for the round. Speakers for each topic shall be indicated by code only.

   c. Preparation Period: Each contestant shall be given forty-five minutes for preparation. A monitor shall be present during the presentation time to assure that there is no consultation and that only the allowable materials are used.

      After drawing his topic, a contestant may not confer with any one nor may he leave the preparation room without the authorization of the room monitor. The contestant may use an annotated bibliography and consult books, magazines, newspapers and summary notes (not to be construed as a prepared speech outline). Students must provide their own preparatory materials. Notes which are organized during the preparation period are permitted for use during the speech.

4. Special Rules:
   a. In all contests, the competition shall begin forty-five minutes after the first contestant has drawn his topic.
   b. The use of typewriters in the preparation room will not be permitted.

5. Time Limit: The time limit for Extempore Speaking shall be six minutes. A speaker shall not be penalized if he speaks less than six minutes.

6. Standards for Excellence:
   a. In delivery, the speaker should be poised, hold attention, use language effectively, have a pleasing voice, and demonstrate true extempore style.
   b. The information used by the speaker should be accurate, pertinent and demonstrate a thorough knowledge of the topic.
   c. The development of the speech should be such that the speaker does not deviate from his topic, progresses with his topic, and demonstrates a reasonable analysis.
   d. The organization of the speech should be such that the introduction gains attention and previews the main points which are clear and in a logical order, and the conclusion is satisfactory.

IHSA SPEECH CRITIQUE
for
Individual Events, Dramatics and Group Interpretation

Event _____ School Code _____

Contestant(s) _____

Title of Selection _____

_____ _____
(Judge's Name and Signature - Please print/type, then sign) (Rank)

Major Basis for Ranking:

Comments:

Southern Illinois University at Carbondale
Department of Speech, Carbondale, Il. 62901

JUDGE'S BALLOT FOR INTERPRETATION

Reader's Name _____ School _____

Selection _____

Assign an evaluation of the reader's effectiveness in each of these items giving 5 if superior; 4 if excellent; 3 if good; 2 if average; 1 if below average; no response if totally ineffective.

I. INTRODUCTION TO THE LITERATURE (_____)
Did the introduction arouse your interest and attract your attention? Did the introduction prepare you sufficiently for the literature?

II. COMMUNICATIVENESS OF THE READER ..... (_____)
Is he direct? Does he contact the audience? Does the audience respond? Is there spontaneity in the reading?

III. PROJECTION OF THE AUTHOR'S MEANING ..... (_____)
Does the reader clearly reveal the theme or central idea? Does the reader grasp and convey the author's meaning?

IV. PROJECTION OF FEELING ..... (_____)
Is the mood conveyed? Does the reader respond emotionally to the thought and mood of the selection? Does he convey this emotion to the audience? Is there emotional control?

V. VOICE ..... (_____)
Does the reader use tone color, imagery, force, rate, variation, emphasis, pause, inflection, climax effectively? Is his voice pleasing and clear? Does it call too much attention to itself?

VI. BODILY ACTION ..... (_____)
Does the reader manage the manuscript effectively? Does he have bodily responsiveness? Do his bodily movements, gestures and facial expression ado to the meaning and feeling of the selection?

VII. TOTAL EFFECT ..... (_____)
Is the effect of the presentation real and genuine rather than mechanical and artificial?

ADDITIONAL COMMENTS:

TOTAL EFFECTIVENESS RATING: Circle the term which best describes the performance.

SUPERIOR EXCELLENT GOOD AVERAGE BELOW AVERAGE

_____
JUDGE

MINNESOTA--ONE ELEMENT OF COMPOSITION

Coherence: 20 points

The paragraph is the basic unit of composition. It deals with one topic or a single division of the topic. It usually contains a topic sentence, and other sentences which develop the idea contained in the topic sentence. Sentences are ordered and linked in such a way that there are smooth transitions within the paragraph.

A longer composition or theme consists of good paragraphs connected smoothly and skillfully with one another. There is a smooth transition of ideas from one paragraph to another. Indentation, or lack of indentation, is not necessarily a reliable guide as a paragraph marker. The clustering of ideas and use of transitions are more reliable indications of good paragraphs. In assessing coherence, one should look for clustering and transition of ideas and not be misguided by the physical appearance of paragraphs caused by indentation or skipping of lines. The allocation of points is variable and cumulative. For example, if 20 points are to be awarded for coherence in a composition, a reader will be free to award from 0 up to the number indicated for the following optimal characteristics:

0 = Writing cannot be intelligibly interpreted. Ideas and sentences are jumbled.

5 = There is little evidence of cohesion. Ideas are barely connected to each other, but they relate to the topic.

10 = Attempts at cohesion are visible. Details and ideas have been gathered and ordered. Transitions are present but not used well.

15 = Fair attempts at cohesion. Details and ideas have been gathered and ordered. Most transitions are used well, but parts of the composition are still separable.

20 = Good evidence of coherence. Details and sections are ordered, there are smooth transitions between paragraphs, and the composition is bound into a wholeness.

DEVELOPING INTER-RATER RELIABILITY: ISSUES TO CONSIDER

One of the stickiest issues in developing consistency among multiple raters occurs when there are varying opinions or ratings of one performance. The focus of discussion at that point should be, *how can we handle unusual conditions?* The focus should **not** be, *who is right?* Remember our goal is that raters will treat like performances in a like manner. We need to find a universal way to codify performance.

In the training process for our local writing assessment at Homewood-Flossmoor Community High School two kinds of performances initially elicited an unusually high level of disparity in scores: 1) when the teachers encountered a paper which did not follow the writing prompt but which was extremely creative, and 2) when the content was clear and well developed but the spelling and mechanics were fraught with error. As a group we had to reach a consensus on how we would treat those circumstances.

A standard score and a standard deduction are two possible approaches to handling such situations.

Standard Score

A performance which exhibits certain qualities can be defined as a particular standard score. For example, in our writing assessment we could decide as a group that any composition written off-topic, no matter how well written, would receive a maximum score of "three" on a four-point scale. Or, we could have gone further and said all off-topic papers will receive a "one" automatically.

A more precise application of a standard score is used to evaluate dives. A description of the system appears in the National Federation of State High School Association's Guidelines for Judging Diving 1987-88:

GUIDELINES FOR JUDGING DIVING

B. JUDGING GUIDE:

10-9	**Exceptional** -- No visible flaws -- approach, hurdle, height, distance, execution and entry all exceptional.
8 1/2-7 1/2	**Excellent** -- Approach, hurdle, height and distance are unaffected. Execution well defined, entry slightly affected.
7-6	**Good** -- Approach and hurdle unaffected; height and distance acceptable; execution defined, entry controlled but slightly affected.
5 1/2-4 1/2	**Satisfactory** -- Approach and hurdle acceptable; height and distance adequate; execution complete and entry slightly over or under.
4-2 1/2	**Deficient** -- Approach and hurdle affected; height and distance inadequate; execution affected and entry over or under.
2-1/2	**Unsatisfactory** -- Approach and hurdle affected; height and distance inadequate; execution not complete or broken position, entry just being made.
0	**Failed dive.**

Standard Deduction

A standard deduction or penality is a second way of increasing rating consistency. Every time a particular error of performance is demonstrated, points can be deducted for the mistake.

The scoring system outlined in the National Federation of State High School Association's 1985-86 Girls Gymnastics Rule Book for rating performance on the balance beam provides a sample of deductive scoring:

B. EXECUTION -- 4.5

1. Execution and Amplitude (E/A) 4.0

a. Faults

Small faults 0.1
Medium faults ..... 0.2 - 0.3
Serious faults 0.4

b. Execution

(1) Flexion of the arms, legs, body; separation of the legs, incorrect position of the hand, hands; pointing toes.
(2) Loss of balance, heavy landing; movements of the arms, trunk, feet to maintain balance.
(3) Stops; rhythm abrupt or jerky.
(4) Touching the apparatus or the floor.
(5) Faults in coming to a support position; alternate placement or delay of the hands.

c. Amplitude

(1) Lack of larger extended and spacious movement
(2) Failure to complete movement with the fullest range of extension

Isolated movements ..... 0.1 - 0.4
Combinations ..... .0.1 - 0.3

NOTE: Refer to specific sections for faults pertaining to that apparatus.

2. General Impression (GI) 0.5

a. Projection (PROJ)

To be deducted each time it occurs within a routine, up to 0.3

(1) Awkwardness of performance 0.1
(a) Audible expression of strain
(b) Visible signs of tiredness/strain
(2) Poor carriage 0.1
(3) Lack of elegance 0.1
(4) Poor appearance 0.1

b. Excellence of Movement (EXM) ..... 0.1 - 0.2

(1) Lack of a superior/high superior difficulty or a series of superior/high superior difficulties performed to the ultimate in technique and with no deductions 0.1
(2) Lack of an entire routine performed to the ultimate in both technique and execution 0.1

No deduction shall exceed the maximum value of that category.

If the point system is to work well, whether it is a standard score or a standard deduction, it must be consensual or agreed upon by all of the teachers involved before rating begins.

RECORDING PERFORMANCE RATINGS

In the preceding section, you saw a variety of rating systems for various activities. Many of these forms were designed to be handed to the student after a performance. How, though, can a teacher keep track of each individual's performance on all of the required objectives? We suggest three relatively simple alternatives: student profiles, objective profiles, or class roster profiles. All of these forms can be produced in schools and duplicated on cardstock so they can be used to record ratings at the site and filed easily later. The forms could also be color coded by year, course, or objective. Some companies have adapted these kinds of forms for machine scoring.

Student Profiles

A summary sheet is developed for each student which lists all of the performance objectives the student should master that year, in a given course or a specific unit. Columns are established for mastery and the date the objective was assessed. For example, a student profile for archery might look as follows:

Wilson, Barbara Student Name	44069 St. I.D.	
P.E. 104 Course	6A Section	Miller Teacher

Y = Mastered N = Not Mastered Date = Date Observed Blank = Not Observed

UNIT OBJ. -- Archery	Trial One Mastery	Date	Trial Two Mastery	Date
1. Addressing the target	Y N	____	Y N	____
2. Notching the arrow	Y N	____	Y N	____
3. Drawing the string	Y N	____	Y N	____
4. Releasing the string	Y N	____	Y N	____
5. Retrieving arrows	Y N	____	Y N	____
6. Hitting the target	Y N	____	Y N	____

Objective Profiles

A class roster is developed for one objective. The advantage of this system is that you can very quickly see how an entire class is doing on a specific objective. Using one of the archery objectives, this form might look as follows:

Alpha Bet	PE 104	6A	____
Teacher Name	Course	Section	Date

INSTRUCTIONAL OBJECTIVE: HITTING THE TARGET

Explanation: Students were judged to have mastered the objective if they were able to hit the target 50% of the time. All students were given a minimum of 5 and a maximum of 10 shots.

Y = Mastered N = Not Mastered Blank = Not Observed

Adams, Barbara	Y	N
Beach, Charles	Y	N
Coates, Diane	Y	N
Diester, Evan	Y	N
Eich, Fran	Y	N
Fay, George	Y	N
Gage, Heidi	Y	N
Hanson, Ira	Y	N
Ilvich, Judy	Y	N

Class Roster Profiles

The class roster profile lists all of the students in a class and the objectives for a course or unit on one page. The objectives are identified only by number. The numbers relate to curriculum guides or are maintained in a thorough key. Abacus Educational Systems, Inc., Denver, Colorado, had NCI develop a scan sheet which is based on the class roster profile idea for use in schools with their Computer Managed Instruction software. We have simulated a portion of their design here as an example. Please note: the actual forms are available commercially.

CMI 3000
CLASS MASTERY SCAN SHEET

Test ID | Teacher ID | Roster ID

0 0 0 0 0 0
1 1 1 1 1 1
2 2 2 2 2 2
3 3 3 3 3 3
4 4 4 4 4 4
5 5 5 5 5 5
6 6 6 6 6 6
7 7 7 7 7 7
8 8 8 8 8 8
9 9 9 9 9 9

Teacher_____
Course Section_____
School_____
Semester_____

STUDENTS	OBJECTIVES							
	1	2	3	4	5	6	7	8
Adams, Barbara	Y N	Y N	Y N	Y N	Y N	Y N	Y N	Y N
Beach, Charles	Y N	Y N	Y N	Y N	Y N	Y N	Y N	Y N
Coates, Diana	Y N	Y N	Y N	Y N	Y N	Y N	Y N	Y N
Diester, Evan	Y N	Y N	Y N	Y N	Y N	Y N	Y N	Y N
Eich, Fran	Y N	Y N	Y N	Y N	Y N	Y N	Y N	Y N
Fay, George	Y N	Y N	Y N	Y N	Y N	Y N	Y N	Y N
Gage, Heidi	Y N	Y N	Y N	Y N	Y N	Y N	Y N	Y N
Hanson, Ira	Y N	Y N	Y N	Y N	Y N	Y N	Y N	Y N
Ilvich, Judy	Y N	Y N	Y N	Y N	Y N	Y N	Y N	Y N

Key to Archery Objectives

1. Addressing theTarget
2. Notching the Arrow
3. Drawing the String
4. Releasing the String
5. Retrieving Arrows
6. Hitting the Target
7. Written Test on Archery
8. Safety Practices and Procedures

Whether you choose one of these options or another, your decision making should include consideration of:

Sufficiency : How you want to retrieve the information.
Accessibility : With whom (and how) you are required to share the information.
Integration : What systems of recording performance you already have in place and the fit of a "new" one.
Practicability : Can you use your system easily in all of the performance sites?
Clarity : If you are planning to store and maintain the records, are they clear enough that they will provide the necessary information later on?

CHAPTER BIBLIOGRAPHY

BASSETT, RONALD E., WHITTINGTON, NILWON, STATON-SPICER, ANN, "The Basics in Spreaking and Listening for High School Graduates: What Should Be Assessed?" Communication Education, Vol. 27, No. 4, November 1978, Pages 293-303.

BERK, RONALD A., Performance Assessment: Methods and Applications. Baltimore, MD: The Johns Hopkins University Press, l986.

CALIFORNIA STATE DEPARTMENT OF EDUCATION, "California Assessment Program: Scoring Guides for the Direct Writing Assessment--Grade Eight." 1987.

CALIFORNIA STATE DEPARTMENT OF EDUCATION, "California Assessment Program: Student Essays Illustrating the CAP Rhetorical Effectiveness Scoring System--Grade Eight." 1987.

CLEARINGHOUSE FOR APPLIED PERFORMANCE TESTING, NORTHWEST REGIONAL EDUCATIONAL LABORATORY, "Annotated Bibliography on Applied Performance Testing in Writing." October 1980.

COLWELL, RICHARD, The Evaluation of Music Teaching and Learning. Prentice-Hall, Inc., New Jersey, 1970.

ILLINOIS STATE BOARD OF EDUCATION, Write on, Illinois!: A User's Guide to Scoring Student Essays. Vol. II, 1987.

KING, HARRY A., SUMMA AUFSESSER, KATHRYN, "Criterion-Referenced Testing--An Ongoing Process." Journal of Physical Education, Recreation, and Dance, January 1988.

LARSON, CARL E., "Problems in Assessing Functional Communication." Communication Education, Vol. 27, No. 4, November 1978, Pages 304-309.

MEHRENS, WILLIAM A., "National Tests and Local Curriculum: Match or Mismatch?" Educatinal Measurement: Issues and Practices. Fall 1984, Pages 9-15.

OREGON STATE DEPARTMENT OF EDUCATION, "Measuring Performance: Verifying Competencies Through Observation and Judgement." 1977, Ed. 144, 990.

RUBIN, DONALD L., DALY, JOHN, MCCROSKEY, JAMES C., MEAD, NANCY A., "A Review and Critique of Procedures for Assessing Speaking and Listening Skills among Preschool through Grade Twelve Students." Communication Education, Vol 31, October 1982, Pages 285-303.

RUBIN, DONALD L., MEAD, NANCY A., Large Scale Assessment of Oral Communication Skills: Kindergarten Through Grade 12. Speech Communication Association, Annandale, Virginia, 1984.

RUBIN, REBECCA, B., "Assessing Speaking and Listening Competence at the College Level: The Communication Competency Assessment Instrument." Communication Education, Vol. 31, January 1982, Pages 20-32.

RUBIN, REBECCA B., "The Validity of the Communication Competency Assessment Instrument." Communication Monographs, Vol. 52, June 1985, Pages 174-185.

SOLLEDER, MARIAN K., "Evaluation Instruments in Health Education." Association for the Advancement of Health Education, 4th Ed., July 1986.

SPEECH COMMUNICATION ASSOCIATION, "Resources for Assessment in Communication." Annandale, Virginia, 1982.

THURSTONE, LOUIS LEON, The Measurement of Values. Chicago: University of Chicago Press, 1959.

PROJECT NOTES

We have included copies of four forms we developed and used in conjunction with the development and rating of our training videotape. The first form is an individual teacher's rating form, the second a cumulative rating form, the third the system of randomization (based on the table of random numbers) we used to determine the order of performance on the training video tape, and the fourth is the release form students signed allowing us to use their taped performance for training. Feel free to use these as you wish.

In addition, we have provided source documentation for all of the formulas used to calculate reliability.

Performance Rating Form--One Rater

SKILL:_____ JUDGE:_____

RATING SCALE: 1 2 3 4 5 6

1. _____	21. _____	41. _____
2. _____	22. _____	42. _____
3. _____	23. _____	43. _____
4. _____	24. _____	44. _____
5. _____	25. _____	45. _____
6. _____	26. _____	46. _____
7. _____	27. _____	47. _____
8. _____	28. _____	48. _____
9. _____	29. _____	49. _____
10. _____	30. _____	50. _____
11. _____	31. _____	51. _____
12. _____	32. _____	52. _____
13. _____	33. _____	53. _____
14. _____	34. _____	54. _____
15. _____	35. _____	55. _____
16. _____	36. _____	56. _____
17. _____	37. _____	57. _____
18. _____	38. _____	58. _____
19. _____	39. _____	59. _____
20. _____	40. _____	60. _____

Cumulative Performance Rating Form--Three Raters

SKILL:_____

	JUDGE 1	2	3	Total		JUDGE 1	2	3	Total		JUDGE 1	2	3	Total
1.					21.					41.				
2.					22.					42.				
3.					23.					43.				
4.					24.					44.				
5.					25.					45.				
6.					26.					46.				
7.					27.					47.				
8.					28.					48.				
9.					29.					49.				
10.					30.					50.				
11.					31.					51.				
12.					32.					52.				
13.					33.					53.				
14.					34.					54.				
15.					35.					55.				
16.					36.					56.				
17.					37.					57.				
18.					38.					58.				
19.					39.					59.				
20.					40.					60.				

Note: We used the cumulative form to record the individual ratings given. This allows for easier calculation of reliability. The worksheet (p. 53) is based on data we obtained from our dance group using these forms.

ORDER OF PERFORMANCES
Based on a Table of Random Numbers

	1	2	3	4	5	6	7	8	9	10
Softball Throwing	8	9	10	1	4	7	3	6	2	5
Shooting a Lay Up	9	6	8	5	2	3	7	4	1	10
Dancing	7	5	4	8	2	3	9	1	10	6
Acting	6	7	8	10	2	3	9	5	4	1
Speaking	9	3	10	5	1	8	4	2	7	6

Steps: 1. Number each performance consecutively from 1-10 in original list.
2. Transform original number to a new random number.
3. Place performances to match the order. For example, the eighth performance comes first in Row 1.

In recording the skill performances on our vcideotape we used the above chart to randomize the sample. We've included it here in case you need to randomize performances.

RELEASE

NAME_____
(Please Print)

ADDRESS_____

I hereby grant Homewood-Flossmoor High School, or its employees, the right to photograph me and use my picture or other reproduction of my physical likeness, including voice recording and/or property. I waive any right to inspect and/or approve the aforementioned. I understand that it is the purpose of Homewood-Flossmoor High School to use the material in a legitimate manner which is not intended to cause harm or undue embarrassment.

_____ _____
Date Signature (Parent of Minor)

_____ _____
Project Director Witness

Illinois State Board of Education
Model Practices Grant: A Prototype
for Determining Validity and
Reliability

_____
Project Title

References for Formulae

Exhibit/Page No.	Source
1. Chapter 1, p. 27 and Exhibit C, Case 2, p. 48	1. Write on, Illinois!: A User's Guide to Scoring Student Essays. Volume II, Illinois State Board of Education, 1987.
2. Exhibit A, pp. 49-52	2. Ebel, Robert L. and David A. Frisbie. Essentials of Educational Measurement (3rd. ed.). Englewood Cliffs, NJ: Prentice-Hall, 1979, pp. 232-284.
3. Exhibit B, pp. 32-34	3. Hoyt, Cyril J. "Test Reliability Estimated by Analysis of Variance." Psychometrika, 1941, 3:153-160.
4. Exhibit D, pp. 34-36 and Exhibit C, Case 1, pp. 47-48	4. Rulon, P.J. "A Simplified Procedure for Determining the Reliability of a Test by Split-Halves." Harvard Educational Review, 1939, 9:99-103.
5. Exhibit E, pp. 40-41	5. Swezey, Robert W. Individual Performance Assessment: An Approach to Criterion-Referenced Test Development. Reston, VA: Reston Publishing Co., 1981, pp. 145-147.
6. Exhibit F, pp. 41-42 and Chapter 1, pp. 12-13	6. Nitko, Anthony J. Educational Tests and Measurement: An Introduction. Orlando, FL: Harcourt Brace Jovanovich, 1983, pp. 406-407.

ANNOTATED BIBLIOGRAPHY

NOTE: Readers who are not familiar with the theory and practice of the assessment of reliability and validity of tests and ratings can learn a little bit about the background and can follow the cookbook recipes in this manual. They will also be able from the manual to make a quick decision about the adequacy of an assessment instrument. Readers wishing to know more about the background of the assessment of competency-based instruments and the technical fine points of validity and reliability methods can find a great deal of help in the following references. We have carefully selected a few of the best references for these purposes. If we were to limit our recommendation to but one volume, it would be Gronlund's Measurement and Evaluation in Teaching.

American Psychological Association. Joint Technical Standards for Educational and Psychological Testing. Washington, D.C.: APA, 1985.

Although focused on standardized testing, the first two chapters can be adapted to the development and use of competency-based mastery tests in schools and/or school districts. Chapter 1, "Validity," gives background on validity; discusses construct-related evidence, content-related evidence, and criterion-related evidence; and gives 25 Standards for validity. Chapter 2, "Reliability and Errors of Measurement," gives background on reliability and gives 12 Standards for reliability.
Even though the material in these two chapters is brief, it is important for professional educators to know that this document exists and that it represents a consensus of three professional organizations whose collective memberships include almost all of the educational and psychological experts in testing and measurement in the U.S.A. In addition to the American Psychological Association, the other two educational associations are American Educational Research Association and National Council on Measurement in Education.

Bork, Ronald A. (Ed.) Performance Assessment: Methods and Applications. Baltimore, MD: The Johns Hopkins University Press, 1986.

Two chapters have excellent sections which are very relevant and up-to-date on performance assessment. More specifically, see Chapter 12, "Personnel Evaluation" by Robert M. Guion (Rater Training, p.356) and Chapter 18, "Writing Skills Assessment," by Edys S. Quellmalz ("Establishing and Maintaining Judges' Ratings,", pp. 500-501) and ("Documenting the Technical Quality of the Assessment System," pp. 501-503).

Cronbach, Lee J. Essentials of Psychological Testing (4th ed.). New York: Harper and Row, 1984.

Chapter 5, "Validation," describes the need for scrutiny of tests, overview of methods of inquiry, validation against a criterion, content validation, and construct validation.
Chapter 6, "How to Judge Tests," describes conceptions of error of measurement, estimating how large errors are, and interpreting findings.
Chapter 15, "Judgments and Systematic Observations," has a major section, "Ratings by Supervisors and Professional Observers," (pp.508-520) which describes the following relevant topics: sources of error, improvement of ratings, scale formats, and scoreable report forms.

Ebel, Robert L. and David A. Frisbie. Essentials of Educational Measurement (4th ed.). Englewood Cliffs, NJ: Prentice-Hall, 1986.

Chapter 5, "The Reliability of Educational Tests," describes the importance of definitions of test reliability, methods of estimating reliability, reliability and errors of measurement, the reliability of content-referenced tests, and factors influencing test reliability.
Chapter 6, "The Validity of Educational Tests," describes the meaning of validity, evidence used to support validity, and questions about validity.

Gronlund, Norman E. Measurement and Evaluation in Teaching (5th ed.). New York: Macmillan, 1985.

Chapter 3, "Validity," describes nature of validity, approaches to test validation, content-related evidence, criterion-related evidence, construct-related evidence, validity of criterion-referenced mastery tests, and factors influencing validity.
Chapter 4, "Reliability and Other Desired Characteristics," describes nature of reliability, determining reliability by correlation methods, standard error of measurement, factors influencing reliability measures, reliability of criterion-referenced mastery tests, how high should reliability be?, and usability.
Chapter 10, "Assembling, Administering, and Appraising Classroom Tests," has a number of sections which could be helpful in revising tests which you have found inadequate in reliability and validity. Especially helpful is the section, "Appraising Classroom Tests" (pp.243-245) with Table 10.1, "Checklist for Evaluating the Classroom Test"; and the section, "Item Analysis and Criterion-Referenced Mastery Tests" (pp.254-257).
Chapter 15, "Evaluating Learning and Development: Observational Techniques," has several sections which might be useful in assessing competency-based achievement. Especially relevant is "Rating Scales" (pp.390-400).
If we were to limit our recommendations to one volume, it would be this book. Gronlund is a bit more easily read, speaks directly to teachers, and is authoritative.

Linn, Robert L. "Instrument Quality Indices in Competency-Based Programs." In Bunda, Mary Anne and James R. Sanders (Eds.). Practices and Problems in Competency-Based Education. Washington, D.C.: National Council on Measurement in Education, 1979.

One of the top leaders in the theory and practice of assessing the quality of competency-based evaluation instruments has shared with us, comprehensively, a history of attempts to assess reliability and validity for competency-based instruments.
His section on reliability includes advantages and disadvantages of various indices and how to cope with the assumption of variability of test scores.
The section on validity is concerned primarily with content validity (most useful to us with competency-based tests), and it thoroughly discusses the rationale for methods of assessing content validity. The secondary roles of criterion-related and construct validations are also discussed.

Nitko, Anthony J. Educational Tests and Measurement: An Introduction. Orlando, FL: Harcourt Brace Jovanovich, 1983.

Chapter 15, "Reliability of Test Scores," describes nature of the concept of reliability, types of reliability coefficients, standard error of measurement, additional factors affecting interpretation of reliability information, and indices of decision consistency.
Chapter 16, "Validity of Test Scores," describes general nature of validity questions, content validity, construct validity, criterion-oriented validity, relating the aspects of validity in order to meaningfully interpret test scores.
Chapter 17, "Criterion-Referenced Interpretations," describes general nature of criterion-referencing, varieties of criterion-references, tests referenced to well-defined and ordered domains, tests referenced to well-defined and unordered domains, concepts that are not equvalent to criterion-referencing, concepts closely related to criterion-referenced testing, reliability and validity of criterion-referenced tests, and establishing cut-off scores for mastery decisions.

Popham, W. James. Modern Educational Measurement. Englewood Cliffs, NJ: Prentice-Hall, 1981.

Chapter 5, "Establishing the Validity of Educational Tests," describes content validity, criterion-related validities, and construct validity.
Chapter 6, "Reliability," describes stability, equivalence, equivalence and stability (combined), internal consistency, the standard error of measurement, and illustrations of published reliability data.

Sax, Gilbert. Principles of Educational and Psychological Measurement and Evaluation (2d ed.). Belmont, CA: Wadsworth Publishing Co., 1980.

Chapter 6, "Ratings and the Reduction of Observational Error," describes the nature of observation, ratings, process and product scales, ranking procedures, improving ratings, and analyzing group behavior.
Chapter 9, "The Reliability of Measurements," describes obtained, true, and error scores; a technical definition of reliability; practical methods of estimating reliability; conditions affecting reliability coefficients; reliability of criterion-referenced tests; the standard error of measurement; the reliability of individual and group scores; the reliability of difference scores; scores of error; and error components of different reliability coefficients.
Chapter 10, "The Validity of Measurements," describes meaning and importance of validity, empirical methods of determining validity: content validity, criterion-related validity, construct validity, factors affecting predictive validity coefficients, cross validation, criterion contamination, face validity, and validity and reasons for using tests.

Swezey, Robert W. Individual Performance Assessment: An Approach to Criterion-Referenced Test Development. Reston, VA: Reston Publishing Co., 1981.

Chapter 7, "Establishing the Reliability and Validity of Criterion-Referenced Tests, includes the topics, advocation of test-retest reliability employing the phi coefficient; three methods of establishing validity (content, concurrent and predictive) of which content validity is recommended as most appropriate for criterion-referenced tests; use of concurrent and predictive test validation when content validity has not been demonstrated; and a discussion of how to deal with tests having low reliability or validity by modifying a criterion-referenced test in order to increase its validity and/or reliability.

SPECIAL THANKS

For supporting this project sufficiently so it could be completed in less than a year

The School Board of District 233
Edward J. Rachford, Ed.D., Superintendent
Donald R. Stoegbauer, Business Manager
Charles W. Smith, Principal

For prereading "Making Sure: Validity and Reliability in Local Testing"

Neil Chance, Principal, and the teachers
Churchill Elementary in Homewood, Illinois
Richard Dervin, Principal, and the teachers
James Hart Junior High in Homewood, Illinois
Charles W. Smith, Principal, and the teachers and administrators
Homewood-Flossmoor Community High School in Flossmoor, Illinois

For developing the assignments and rating criteria and selecting performances

Throwing a Ball
John Wrenn, Chair
Susan Kamba
Horace Howard

Shooting a Lay-up
Faye Blake, Chair
Val Middleton
Tom Skill

Dancing: One Combination
Debbie Hinchcliffe, Chair
Lynn Frank
Elizabeth Mahar

Acting: Characterizing Old Age
Tom Sweeney, Chair
Lyle Wathier
Melissa Chadwick
David Boettcher

Speaking: Delivering an Original Speech
Megan Tipton, Chair
Barry Kautz
Candace DuBois

For their serving as the project Video Production Team

Lorelei Jones, Megan Tipton
student assistants:
Joe Bandy, Marc Chambers, Chuck Garfien, Josh Gerber, Nitika Gondal, Mike Linde, Larry Rogove

For patiently entering, arranging, and producing text: Materials Production Team

Cheryl Kozlow, Gloria Schlinger, Carolyn Wig, Marlene Drucker and Ruth Ruskowsky

For providing resources and for answering our every inquiry

Illinois State Board of Education, Department of School Improvement Services

Program Development & Delivery Section
Ray Schaljo
Ann Heelen

Curriculum Improvement Section
Rob Sampson
Roberta Volkmann
Mary Jo Leeds
Susan Richardson

Student Assessment Section
Tom Kerins
Merv Brennan
Judy Dawson
Eunice Greer
John Craig
Carmen Chapman

Graphics designed by LORELEI JONES.

District 233

Homewood-Flossmoor Community High School

Dr. Edward J. Rachford, Superintendent

999 Kedzie Avenue

Flossmoor, Illinois 60422

Phone: (312) 799-3000